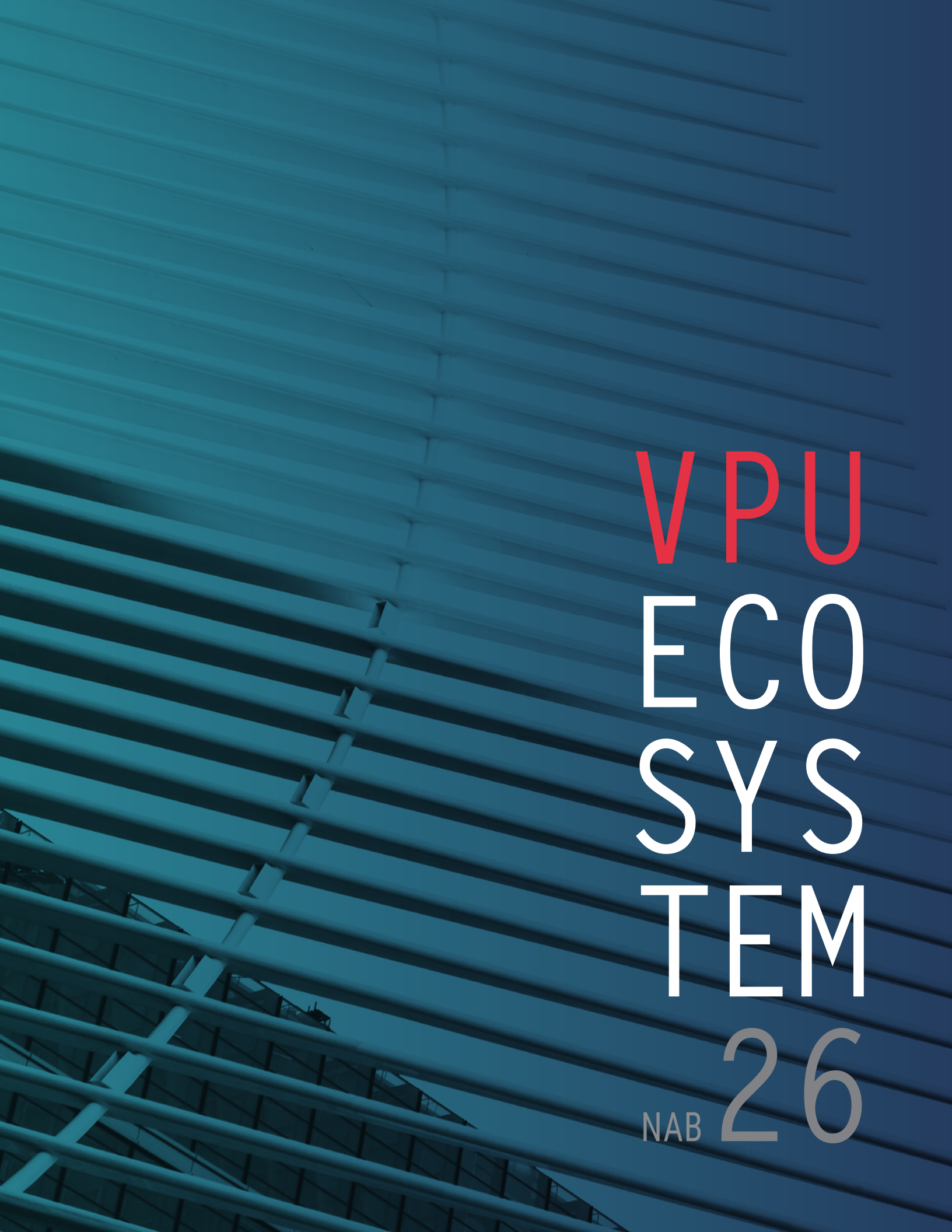


VPU ECOSYSTEM

NAB 26

THE ERA THAT JUST ENDED

~~Borrowed hardware.~~
~~General-purpose compute.~~
~~Software that compensates.~~

THE ERA THAT HAS ARRIVED

Purpose-built silicon.
Deterministic infrastructure.
Economics that compound.

What follows is the proof.

NETINT
VPU ECOSYSTEM

The VPU did not arrive as a product pitch.

It arrived as an answer to a physics 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

NETINT TECHNOLOGIES

Table of Contents

THE NEW ECONOMICS OF VIDEO INFRASTRUCTURE

Why efficiency, not flexibility, is becoming the primary design constraint for modern video systems

THE SHIFT

~~FLEXIBILITY~~ *was the goal.*

EFFICIENCY **is now the constraint.**

SILICON DESIGNED FOR VIDEO

CPU
General computation

GPU
Parallel graphics & AI tasks

VPU
Purpose-built ONLY for video encoding. Nothing else.

THE SHARED VIDEO PIPELINE

Ingest → Transcode & Process → Package & Origin → Deliver → Secure & Optimize

The VPU Ecosystem is not a vendor story. It is a design philosophy.

Escape CPU encoding and cut OPEX 90%

NETINT VPU ECOSYSTEM

NETINT VPU ECOSYSTEM

INFRASTRUCTURE STRATEGY

The New Economics of Video Infrastructure

Why Efficiency, Not Flexibility, is Becoming the Primary Design Constraint for Modern Video Systems.

For two decades, the video industry built its infrastructure on borrowed hardware — compute designed for other jobs, adapted to carry the weight of global streaming. CPUs designed for enterprise workloads. GPUs designed for graphics and AI. Software encoders patched on top, optimizing around limitations that were never meant to be solved. It worked well enough during the growth phase. It is no longer working well enough at scale.

But something has shifted. The dominant pressures facing video operations today are no longer about what a system can do. They are about what it costs to keep doing it. Concurrency scales faster than audiences. Adaptive bitrate ladders expand faster than encoding efficiency improves. Power, density, and predictability now matter as much as raw throughput.

This landscape gave rise to the Video Processing Unit, or VPU: purpose-built silicon designed specifically for video workloads. And the implications of VPU adoption extend far beyond a hardware swap. They are reshaping how teams plan, build, and operate video systems at every layer of the stack.

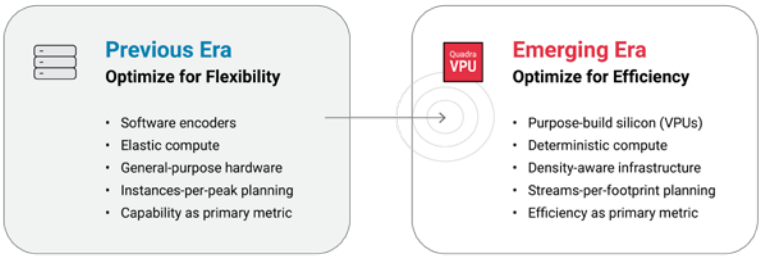


Exhibit 1: The video industry's optimization target is migrating from flexibility to efficiency.

The Constraint Has Moved

To understand why VPUs matter, it helps to understand what changed. In the early era of streaming, the challenge was capability: could your system encode, package, and deliver video at all? Hardware was expensive, but workloads were modest. Software encoders running on commodity servers were good enough.

Today, the problem is different. Platforms are struggling to process video economically at sustained scale. A single live sporting event can require thousands of concurrent transcode sessions. Each session demands multiple output renditions across an expanding set of resolutions, codecs, and device profiles. The math compounds quickly: more streams, more renditions, more power, more cooling, more cost. The flexibility that once made software encoders attractive is now a liability when the priority is predictable, efficient throughput at density.

KEY INSIGHT

The constraint in video infrastructure has moved. It's no longer about what your system can do. It's about what it costs to keep doing it, at the scale the market now demands.

What a VPU Actually Is, and What It Is Not

A Video Processing Unit is purpose-built silicon designed specifically for video encoding and media processing workloads. Unlike a CPU, which is optimized for general computation, or a GPU, which is optimized for parallel graphics and AI tasks, a VPU is architected to do one thing exceptionally well: transcode video streams efficiently and deterministically. That distinction matters.

A VPU is purpose-built silicon architected for one function: transcoding video streams efficiently, deterministically, and at density. That specificity is not a limitation. It is the point. In infrastructure design, the most powerful tools are not the most flexible ones — they are the ones built precisely for the job at hand. VPUs change the cost curve, the power equation, and the operational predictability of video transcoding in ways that general-purpose compute cannot replicate. That is not a product claim. It is a physics argument.

In the VPU Ecosystem framework, VPUs are treated not as a product feature but as an infrastructure primitive: an architectural lever that reshapes how teams think about capacity, power, and operational predictability.

A Shared Pipeline, Different Pressures

The VPU Ecosystem organizes the modern video stack around a shared pipeline: Ingest, Transcode and Process, Package and Origin, Deliver, and Secure and Optimize. Every company in the ecosystem operates somewhere along this chain. Some design and supply infrastructure. Others operate platforms at scale. Still others orchestrate, integrate, or govern how video processing is deployed within existing systems.

Not every partner in the ecosystem directly deploys VPUs. But all of them are influenced by the constraints VPUs introduce, particularly around density, power efficiency, predictability, and deployment friction. This enables the partner to compete in a whole new way by offering higher quality, lower cost, and better value than their competitors.

This is the critical insight: the value of the VPU Ecosystem lies in how efficiency-driven constraints ripple across every stage of the pipeline, reshaping architectural and operational decisions. VPUs are driving improved end user experiences, while opening new monetization opportunities for platforms, streaming services, and the technology providers powering these systems.

Exhibit 2: The shared video pipeline model, a common reference framework across the VPU Ecosystem.

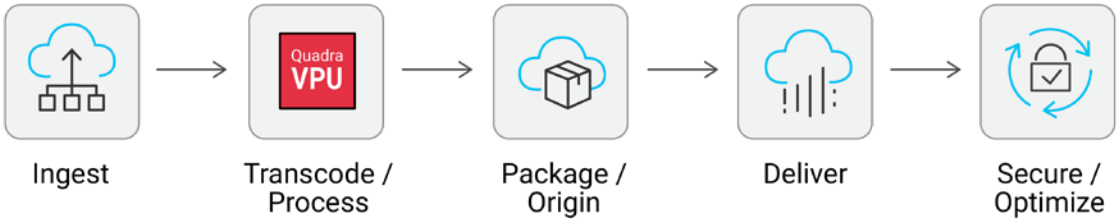


Exhibit 3: Three system-level effects observed consistently when VPUs are introduced into video workflows.

Three Effects That Change Everything

General-purpose compute doesn't fail dramatically. It degrades quietly — and the degradation accelerates at exactly the moment scale increases. VPU introduction reverses three of those degradation patterns at once.

Across environments (cloud, on-prem, hybrid), introducing VPUs consistently changes three system properties. Understanding these effects is essential for any leader evaluating video infrastructure strategy.

Density becomes a planning unit. When VPUs enter the stack, teams stop thinking in instances-per-peak and start thinking in streams-per-footprint. This is a subtle profound shift. It means infrastructure planning moves from reactive scaling to proactive capacity design, a change that affects procurement, facility planning, and long-term capital allocation.

Power becomes a visible constraint. Energy consumption per stream has historically been an afterthought, buried in operational budgets and ignored in system design. VPUs make it a first-order variable. Energy per stream now influences facility design, sustainability reporting, and long-term cost modeling, even in cloud environments where power costs are embedded in compute pricing.

Predictability improves. General-purpose compute is elastic by design, which introduces variability. VPU-based transcoding trades some of that elasticity for stability. The result is more deterministic performance, simpler capacity planning, and fewer surprises during peak load events. For operations teams, this is transformational.

THE BIGGER PICTURE

The VPU Ecosystem is the infrastructure model the industry is converging on, because the economics of video at scale demanded it. Efficiency, density, and predictability are not features. The Ecosystem exists to make that architecture deployable, at every layer of the stack.

What This Means for Decision-Makers

The shift toward efficiency-first video infrastructure is already underway. The organizations that recognize it earliest will have the clearest advantage, not because they adopted a specific technology, but because they redesigned their systems around the right constraints.

For platform engineers, this means evaluating transcoding architecture not just on throughput but on density and power per stream. For operations leaders, it means building capacity plans around deterministic baselines rather than elastic headroom. For finance and strategy teams, it means modeling infrastructure costs with sustainability and energy efficiency as explicit variables, not opaque line items.

The essays and case studies that follow in this series each examine one specific dimension of this ecosystem. They explore how the pressures of economics, latency, orchestration, production reality, deployment risk, and operational scale play out differently depending on where a company sits in the pipeline. Read together, they form a comprehensive picture of an industry in transition, and a practical framework for navigating it.

Advantech shows how the server architecture surrounding a VPU, PCIe topology, power delivery, and thermal design, is the factor that determines whether silicon efficiency becomes real-world, sustained infrastructure performance. *Page 9*

Akamai Cloud will explain how pairing VPU-accelerated transcoding with their Distributed Cloud creates a highly cost-effective and energy-efficient architecture for global video delivery. *Page 15*

Arcadian explains how the predictable behavior and bounded power consumption of VPUs provide the necessary operational margin and reliability to survive unpredictable, real-world live production constraints. *Page 20*

Cires21 details how to fairly measure and validate efficiency, quality, and scale in heterogeneous video environments by focusing on system-level behavior and stability under sustained loads. *Page 27*

Dell Technologies shows how integrating efficient VPUs into validated server platforms translates silicon-level power and density gains into deployable, scalable, and supportable datacenter level video infrastructure. *Page 32*

i3D.net shares how the efficiency and density of VPUs enable latency-sensitive video processing to be geographically distributed closer to users without multiplying infrastructure costs. *Page 53*

Misao Network presents a case study showing the combination of low-latency WebRTC video transport with VPUs enables the delivery of broadcast-quality 8K video for real-time live environments. *Page 59*

Scalstrm shows how advanced orchestration and control planes are essential for managing complex, VPU-accelerated video pipelines to turn raw hardware efficiency into operational reliability. *Page 65*

V-Nova shares how MPEG-5 LCEVC technology enable video streaming platforms to deliver superior visual quality at lower bitrates across a wide range of video streaming and entertainment services architectures. *Page 71*

THIS ARTICLE IS THE FIRST IN THE VPU ECOSYSTEM SERIES, A COLLECTION OF FEATURE ARTICLES AND CASE STUDIES EXAMINING HOW EFFICIENCY-DRIVEN DESIGN DECISIONS ARE RESHAPING MODERN VIDEO INFRASTRUCTURE. EACH PIECE FOCUSES ON ONE OPERATIONAL PRESSURE AND IS INTENDED TO BE READ AS A COMPLEMENTARY PERSPECTIVE WITHIN THE BROADER ECOSYSTEM FRAMEWORK.

Flexible VPU Platforms for Live Streaming Workflows

Studio → Edge → OB Truck → Data Center → Cloud

ADVANTECH

HARDWARE PLATFORM

The Architecture Beneath the Accelerator

Why System Design Determines Whether Video Processing Units Deliver on Their Promise.

The video industry has spent the past decade pursuing a deceptively simple goal: process more streams with less hardware. The emergence of Video Processing Units, or VPUs, has brought that goal within reach. Purpose-built ASIC silicon like NETINT's Quadra can encode hundreds of HD channels in a single device, consuming a fraction of the power that CPU-based software encoding requires. But efficiency measured on a lab bench and efficiency sustained in a production environment are not the same thing.

The server that hosts an accelerator determines whether its theoretical performance translates into reliable, sustained throughput. PCIe topology governs how data moves between the host processor and the accelerator. Power architecture dictates whether multiple devices can operate simultaneously at full load. Thermal design determines whether performance holds steady over hours and days of continuous operation. Mechanical layout decides whether a failed component can be replaced without taking the entire system offline.

For organizations deploying accelerator-dense video infrastructure at scale, system design is the factor that separates a successful deployment from an expensive disappointment.

When Silicon Gets Faster, Systems Must Get Smarter

Traditional video infrastructure evolved around general-purpose CPUs, and later GPUs. The servers that hosted these processors were designed for balanced workloads where compute, storage, and networking shared relatively even resource profiles. A standard 1RU or 2RU server could handle video encoding alongside other tasks without requiring special attention to any single subsystem.

Accelerator-based workloads behave differently. Video processing places sustained, predictable pressure on specific system resources: PCIe bandwidth for moving uncompressed frames between host and device, power delivery for maintaining multiple accelerators under continuous load, cooling pathways for dissipating concentrated heat from devices that never idle, and storage throughput for buffering high-bitrate content.

When these resources are not aligned with workload behavior, system-level bottlenecks appear long before the silicon reaches its limits. An accelerator capable of encoding 80 streams might deliver only 50 if the PCIe bus cannot feed it data fast enough. A card rated for 75 watts might throttle to 60 if the chassis airflow cannot keep it within operating temperature. The result is stranded capacity: hardware that was purchased for its peak specification but can only sustain a fraction of it.

KEY INSIGHT

The performance ceiling of an accelerator-dense system is not set by the accelerator. It is set by the weakest link in the chain of system resources that support it. Purpose-built platforms are designed to ensure that no single resource becomes that bottleneck.

Density Changes the Planning Model

The efficiency gains from VPUs create a natural incentive to increase density. If one accelerator card can replace an entire CPU-based encoding server, the logical next step is to pack multiple cards into a single chassis and consolidate even further. NETINT's Quadra family of VPU's are available in PCIe, U.2, and M.2 form factors, enabling configurations that can scale from two devices to twelve or more on a single platform.

Increasing density introduces compound constraints. Each additional accelerator draws power from the same mother-board, generates heat within the same airflow envelope, and competes for bandwidth on the same PCIe bus. Systems designed for moderate accelerator counts may struggle when those devices operate continuously under sustained load.

Effective system design must anticipate not just peak throughput, but sustained concurrency. A platform that benchmarks well with four accelerators running short test sequences may reveal thermal or power limitations when those same four accelerators encode live streams around the clock. The planning model shifts from 'how many devices fit?' to 'how many devices can operate at full capacity simultaneously, indefinitely?

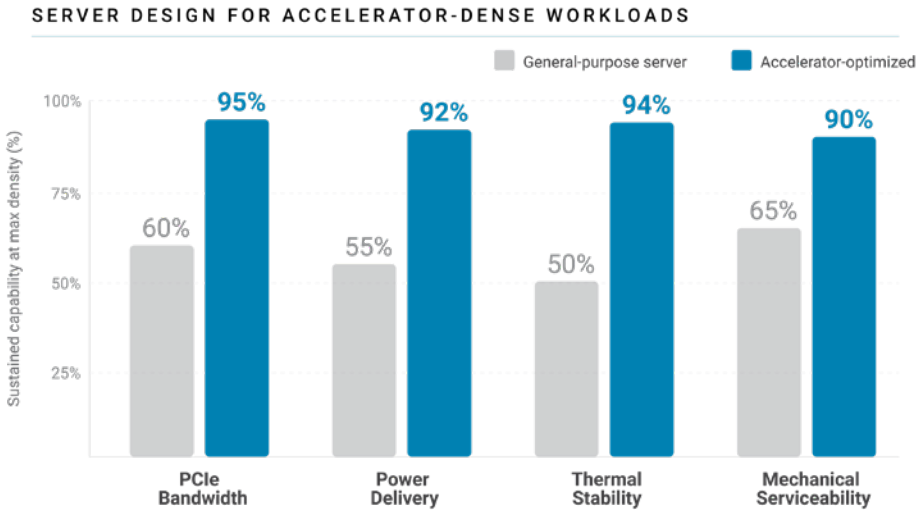


Exhibit 1: At high accelerator density, purpose-built platforms sustain significantly more capability across all critical system resources.

The Five Pillars of Accelerator-Ready System Design

The system-level considerations that determine real-world accelerator performance fall into five interconnected categories. Each affects the others, and a weakness in any one can limit the entire platform.

SYSTEM DESIGN FOR ACCELERATOR-DENSE VIDEO PLATFORMS

Design Area	Key Considerations	Impact on Accelerator Workloads
PCIe Topology	Lane distribution, switch hierarchy, bus contention	Throughput ceiling for multi-accelerator configurations
Power Architecture	Per-slot wattage, VRM placement, load balancing	Sustained operation without brownouts or throttling
Thermal Design	Separated fan zones, directed airflow, ambient tolerance	Consistent performance over hours and days
Mechanical Layout	Hot-swap access, cable routing, accelerator spacing	Field serviceability without full system downtime
Chassis Depth	Short-depth vs. full-depth rack compatibility	Edge and telecom deployment flexibility

FROM SILICON EFFICIENCY TO DEPLOYABLE INFRASTRUCTURE

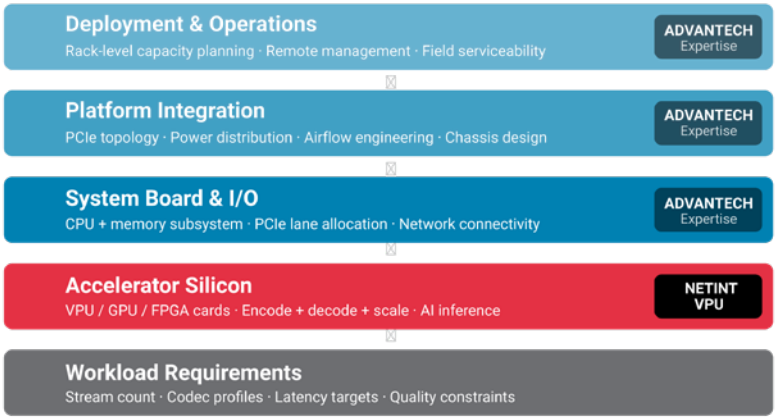


Exhibit 2: Purpose-built platforms bridge multiple layers between silicon efficiency and deployable infrastructure.

The Hidden Importance of PCIe Topology

Of all the system design factors, PCIe topology may be the least visible and the most consequential. Accelerator-based video systems depend on PCIe communication between the host processor, the accelerator devices, and storage. Every uncompressed video frame that enters an encoder and every compressed bitstream that exits it travels across this bus.

When multiple accelerators share limited PCIe bandwidth, contention can degrade throughput even if individual devices remain underutilized. In high-density systems, the details matter: lane distribution across slots, switch placement on the motherboard, and the hierarchy of the PCIe bus all affect achievable performance. A system with four PCIe x16 slots might deliver excellent bandwidth to each device individually, but if those slots share a common root complex or pass through an oversubscribed switch, concurrent operation may tell a different story.

Well-designed platforms treat PCIe connectivity as a first-order system resource rather than an afterthought. Advantech's VEGA-7000 series, for example, provides four PCIe expansion slots in a 1U form factor, with lane allocation specifically engineered for sustained accelerator throughput. The goal is predictable, non-contending bandwidth for every device in the system, not just the first one.

Thermal Design Determines Sustained Performance

Accelerator workloads differ fundamentally from burst-oriented compute tasks. Video processing often runs continuously for hours, days, or weeks, particularly in live streaming and broadcast environments. Under these conditions, thermal stability becomes critical.

Poor airflow design leads to device throttling, reduced encoding throughput, and increased hardware failure rates. The challenge intensifies with density: more accelerators in the same chassis means more concentrated heat, and the thermal output of upstream devices can preheat air reaching downstream ones.

Advantech's approach to this challenge includes separated fan zones for CPU and accelerator areas, smart fan control through IPMI that adjusts speeds based on individual component temperatures, and directed airflow paths that ensure each device receives cooling independently of its neighbors. The SKY-6000 series GPU servers, for instance, use separate air tunnels for CPU and accelerator areas to prevent thermal cross-contamination. This is not a minor engineering detail. It is the difference between a system that maintains rated performance at hour one and a system that still maintains it at hour ten thousand.

THE BIGGER PICTURE

Thermal throttling is invisible to operators until it manifests as degraded stream quality or dropped frames. A well-designed platform eliminates the conditions that cause throttling in the first place, making performance predictable rather than aspirational. For 24/7 video operations, this predictability is the foundation of service reliability.

Edge Deployments Amplify Every Design Decision

Not all video processing occurs in large, climate-controlled data centers. Edge deployments at regional sites, telecom facilities, or production environments introduce additional constraints: limited power budgets, restricted physical space, constrained cooling capacity, and environmental variability. A platform that performs flawlessly in a temperature-controlled server room may behave very differently in a telecom central office or a broadcast production truck.

These environments are precisely where accelerator density matters most. Edge sites often lack the space and power for multiple large servers, which makes the ability to consolidate more processing into fewer, smaller platforms essential. Advantech's portfolio addresses this directly, offering both short-depth 1U servers like the VEGA-7030 for space-constrained deployments and full-depth multi-accelerator platforms for regional data centers. The VEGA-7030, designed specifically for video workloads, fits a short-depth rack footprint while still accommodating four PCIe accelerator cards.

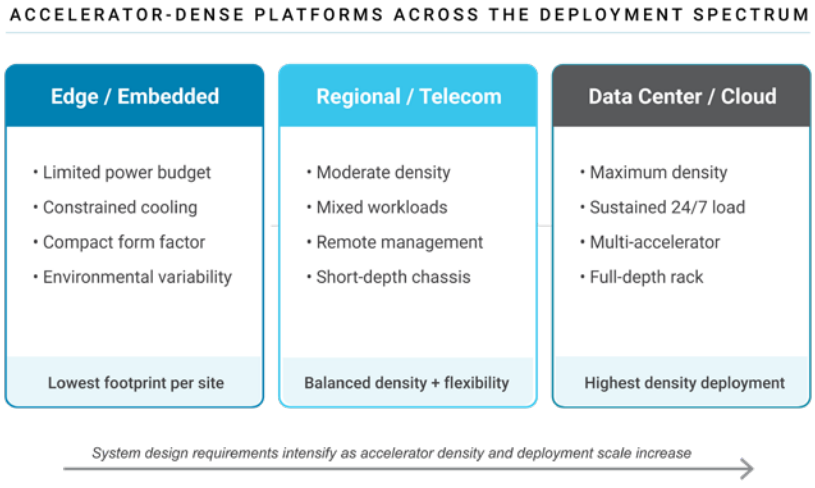


Exhibit 3: System design requirements intensify as deployments move from controlled data centers to constrained edge environments.

The primary benefit of specialized video hardware is predictability. Encoding throughput, power consumption, and performance remain stable under sustained workloads. A VPU server that processes 320 streams or more at a given quality level will continue to process those 320 streams at that quality level for the duration of the workload. This is fundamentally different from CPU-based encoding, where performance can vary with system load, thermal conditions, and competing processes.

Effective platforms therefore extend the predictability of VPUs into the full system. The result is infrastructure where capacity planning can be based on actual, sustained performance rather than theoretical peak specifications.

The Strategic Calculus

Video infrastructure is increasingly shaped by specialized hardware. VPUs dramatically improve the efficiency of video processing, but their real impact depends on how they are integrated into systems. Server architecture, device density, and thermal stability ultimately determine whether efficient silicon becomes deployable infrastructure.

Within the VPU ecosystem, Advantech represents the system design layer: the point where accelerator efficiency becomes operational reality. Rather than treating accelerators as optional add-ons to general-purpose servers, Advantech designs platforms where accelerators are central to the system's purpose. Optimized PCIe topology ensures predictable bandwidth. Robust power architecture supports sustained multi-device operation. Intelligent thermal management maintains performance over continuous, long-duration workloads.

For platform architects and infrastructure engineers evaluating their next generation of video processing systems, the question is not simply which accelerator to choose. It is whether the system that surrounds it can deliver on the accelerator's promise, consistently, at scale, in the environments where it actually needs to operate.

That question is answered not by the silicon, but by the architecture beneath it.

THIS ARTICLE IS PART OF THE VPU ECOSYSTEM SERIES EXPLORING HOW PURPOSE-BUILT SILICON IS RESHAPING VIDEO INFRASTRUCTURE ACROSS THE TECHNOLOGY STACK. EACH ARTICLE EXAMINES A DIFFERENT ARCHITECTURAL PERSPECTIVE THROUGH THE LENS OF ONE PARTNER WITHIN THE ECOSYSTEM.


Akamai Cloud
NAB SHOW
Accelerated Compute
NETINT VPUs deployed globally
and right-sized for efficiency
Lower your cloud costs
for media-intensive tasks

VIDEO INFRASTRUCTURE

When Economics of Video Break: A Framework for Escaping the 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.

The Four Cost Levers That Break Video Infrastructure

To understand why traditional approaches fail, it helps to identify the four levers that drive infrastructure costs in video. Each one is significant on its own. Together, they create a compounding effect that makes linear scaling unsustainable.

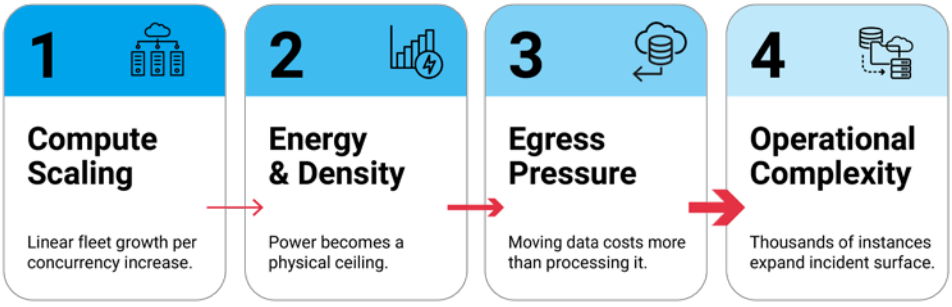


Exhibit 1: The four compounding cost levers in video infrastructure.

1. The first lever is compute scaling.
CPU-based encoding scales linearly with fleet size. Double the concurrency, double the instances. For a high-profile live event, this means provisioning thousands of general-purpose machines for a task that lasts hours.

2. The second lever is energy and density.
Power consumption is no longer an abstract line item. It's becoming a physical ceiling. Even in cloud environments, this manifests as regional capacity limits and pricing volatility that create unpredictable cost spikes.

3. The third lever is egress pressure.
In many architectures, the cost of moving transcoded video from the compute layer to the delivery layer exceeds the cost of the transcoding itself. This is the hidden tax that most ROI models underestimate.

4. The fourth lever is operational complexity.
Managing a fleet of general-purpose instances for a specialized media task expands what engineers call the "incident surface area." More instances mean more failure points, more monitoring overhead, and more DevOps hours spent on infrastructure rather than product.

KEY INSIGHT

These four levers do not operate independently. A spike in concurrency (Lever 1) increases power draw (Lever 2), generates more egress traffic (Lever 3), and requires more fleet management (Lever 4).

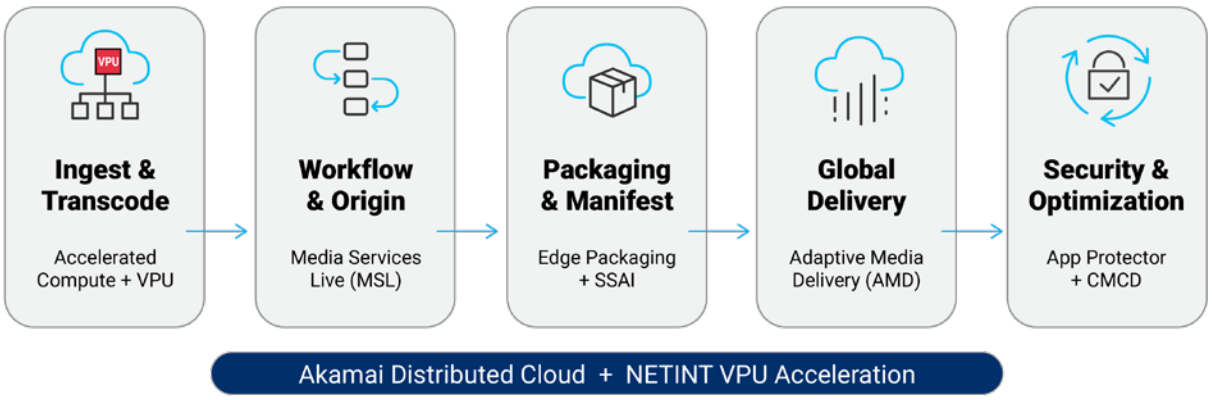
The compounding nature of this cycle is what makes incremental optimization insufficient.

A 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."

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



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.

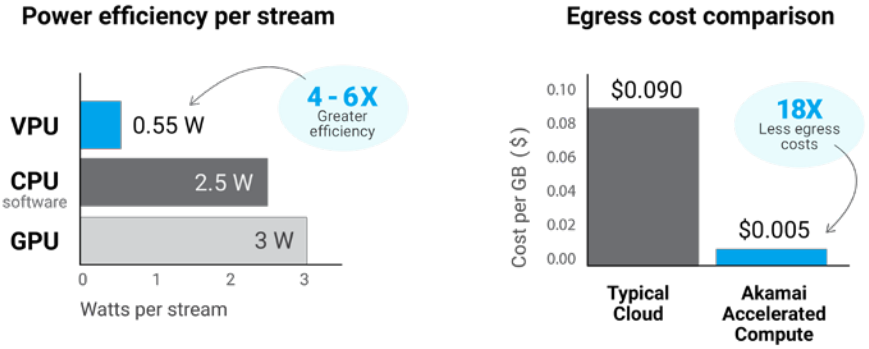


Exhibit 3: Power efficiency and egress cost comparison *(source: published Cires21 + NETINT benchmarks).*

The 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 to 3.6 watts per stream, a VPU sustains equivalent quality at approximately 0.4 to 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.

THE BIGGER PICTURE

When egress costs drop by an order of magnitude, the calculus around ABR ladder design changes entirely. Organizations can afford to include higher-bitrate rungs that were previously cost-prohibitive, directly improving the viewer experience without proportional cost increases.

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.

THIS ARTICLE IS PART OF THE VPU ECOSYSTEM SERIES EXPLORING HOW PURPOSE-BUILT SILICON IS RESHAPING VIDEO INFRASTRUCTURE ACROSS THE TECHNOLOGY STACK. EACH ARTICLE EXAMINES A DIFFERENT ARCHITECTURAL PERSPECTIVE THROUGH THE LENS OF ONE PARTNER WITHIN THE ECOSYSTEM.


The Future
of Hardware Acceleration
Discover a smarter way to transcode
by lowering your cloud costs for
media-intensive tasks.
Akamai Cloud

PROFESSIONAL SERVICES

When Theory Meets Production

Building Video Systems That Survive Real-World Conditions.

Every vendor publishes benchmarks. But video infrastructure decisions that matter most are rarely made in the lab. They are made in production environments where content arrives malformed, workloads shift without warning, and the cost of failure compounds every hour. This is the gap between theory and practice, and it is wider than most technology evaluations acknowledge. Benchmark numbers describe what a system can do under ideal conditions. Production reveals what it will do under pressure. The distinction matters because infrastructure teams do not get to choose their conditions. They inherit them.

Arcadian, a Los Angeles-based media services firm founded in 2017, has spent years operating in this space. As a trusted partner to studios including Sony Pictures and platforms like Crunchyroll, Arcadian has processed and delivered over 150,000 hours of video content across its transcoding, QA, and deployment operations. Their perspective on infrastructure is shaped by the unpredictable reality of large-scale media delivery. When they integrated NETINT VPU technology into production workflows, the results validated what their operational experience had long suggested: the systems that survive real-world conditions are built around predictability, not peak performance.

Production Surfaces Different Constraints

Lab environments are, by design, controlled. Test clips are clean. Codecs are configured in isolation. Workloads are tuned to demonstrate best-case throughput. This is not a criticism of benchmarking. It is a recognition that benchmarks answer a specific, narrow question: how does this system perform when everything goes right?

Production asks a different question: how does this system behave when everything is messy?

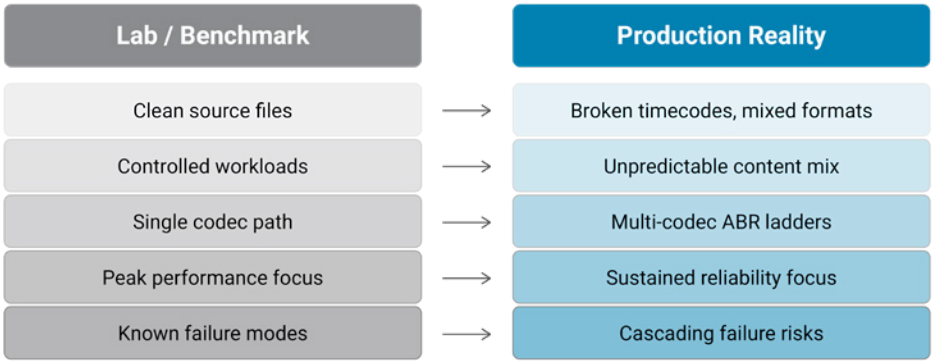


Exhibit 1: What Changes When Systems Leave the Lab.

In Arcadian's experience managing transcoding operations for major studios, the differences between lab and production are stark. Source media arrives with broken time codes, inconsistent frame rates, and incompatible encoding settings. A single content library may contain files spanning decades of production standards, from legacy MPEG-2 assets to modern 4K HDR masters. The transcoding system must handle all of them without manual intervention for each exception.

Workloads are equally unpredictable. A steady queue of standard-definition catalog content can be interrupted by an urgent batch of premium 4K titles requiring Dolby Vision and Dolby Atmos processing. The system that benchmarked beautifully for one scenario may struggle when both arrive simultaneously. Production infrastructure must accommodate this variability as a baseline condition, not as an edge case.

Reliability Is Defined by Predictability

In operational contexts, reliability is not simply uptime. It is the degree to which a system's behavior can be anticipated. A transcoding pipeline that produces different output characteristics on successive runs of the same input is unreliable, even if it never technically fails. Operators need to know that results will be consistent, that throughput will remain stable under varied loads, and that scaling the system will not introduce new failure modes.

Arcadian's engineering and QA teams have observed this principle firsthand. Their hybrid QA model combines automated testing with trained human specialists who serve as the final arbiters of quality. What they found with VPU-based processing was striking: the same input consistently produced the same output, regardless of how many times it was transcoded or how many parallel streams were running.

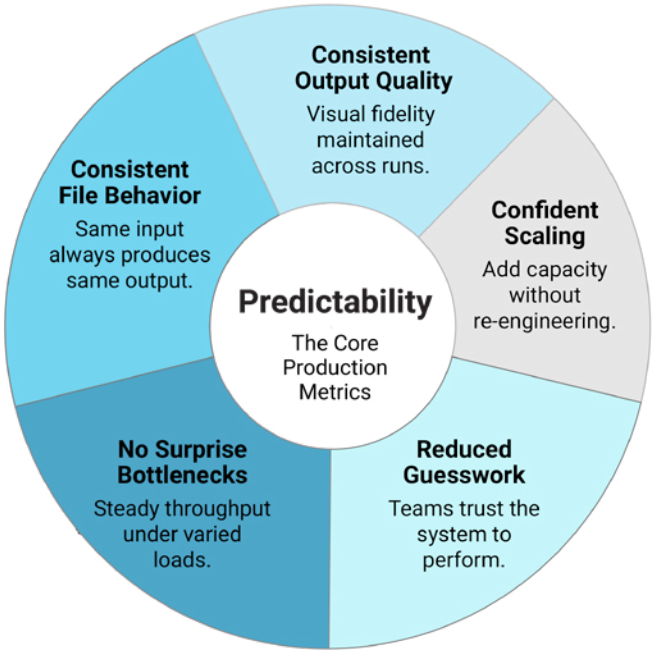


Exhibit 2: Five Dimensions of Production Predictability are issues that matter most to production teams.

This predictability extends across five dimensions that matter to production teams.

- First, consistent file behavior: the same source always yields the same output.
- Second, reduced guesswork: operators trust the system to perform without constant monitoring.
- Third, no surprise bottlenecks: throughput remains steady whether the queue holds ten jobs or ten thousand.
- Fourth, confident scaling: adding capacity is additive, not a re-engineering exercise.
- Fifth, consistent output quality: visual fidelity is maintained across runs, codecs, and resolutions.

For a company processing content destined for platforms like Sony Pictures Core, PlayStation, and BRAVIA, this consistency is not an abstract virtue. It is a contractual requirement. When Arcadian delivers content, the receiving platform expects every asset to meet specification. Predictability at the infrastructure level is what makes that guarantee possible at the business level.

THE BIGGER PICTURE

Efficiency in production is not a single number. It is the compound effect of lower power draw, higher density, reduced cooling overhead, and extended hardware life, measured over months and years of sustained operation rather than minutes of peak performance.

Why Efficiency Matters Differently in Production

Efficiency in a benchmark context typically means operations per second or frames per watt. These are useful metrics, but they do not capture how efficiency functions in a production environment. In production, efficiency is about operational margin: the gap between what a system costs to run and the value it produces over sustained periods.

Consider power consumption. A GPU-based transcoding card drawing 250 to 400 watts handles a known number of streams. A VPU-based solution handling comparable workloads at a fraction of that power draw does not just reduce the electricity bill. It reduces cooling requirements, extends hardware lifespan, and allows denser deployments in existing rack space. Arcadian's deployment of NETINT Smart VPU technology demonstrated precisely this: up to ten times more streams per rack unit compared to CPU-based alternatives, with 80% less power consumption.

The economics compound over time. A single rack that previously held two GPU-based encoding servers can accommodate four or more VPU-equipped systems in the same physical footprint. The colocation costs remain fixed while the encoding capacity multiplies. For an operation like Arcadian's, which manages transcoding at scale for clients including Outdoor Channel Asia and VIVA Entertainment in addition to its Sony partnership, this density advantage translates directly into the ability to serve more clients without proportional infrastructure expansion.

What VPUs Change in Real-World Deployments

The architectural difference between a GPU and a VPU is not simply a matter of silicon design. It manifests in how the hardware integrates into existing infrastructure, how it behaves under production conditions, and how teams interact with it on a daily basis.

Arcadian's operational teams describe VPU deployment as dramatically simpler than GPU alternatives. NETINT's Quadra VPUs use standard PCIe and NVMe form factors, which means they slot into existing server chassis without specialized mounting, cooling, or power configurations. A single server can accommodate four or more VPU cards where it might hold only one or two GPUs. The driver and software layer is equally streamlined: NETINT provides

Docker-ready container images that reduce deployment from days of configuration to hours of integration.

In practice, this simplicity matters more than raw throughput numbers. When Arcadian needs to scale capacity for a large content delivery deadline, the operational question is not whether the hardware is fast enough. It is whether new capacity can be brought online quickly, integrated with existing pipelines reliably, and scaled back when the project completes. VPU architecture supports this operational rhythm because it was designed around production requirements rather than retrofitted from general-purpose compute.

The codec coverage reinforces this production orientation. NETINT Quadra supports H.264, HEVC, and AV1 encoding, which covers the full range of delivery requirements Arcadian encounters across its client base, from legacy catalog content to next-generation streaming formats including 8K HDR. A single hardware platform that handles the complete codec ladder eliminates the need for separate encoding paths for different output requirements.

Failure Modes That Production Exposes

Perhaps the most revealing test of any infrastructure is not how it performs at peak, but how it fails. Production environments generate failure modes that controlled tests rarely surface, and the cost of those failures can escalate rapidly.

Arcadian's teams have documented three categories of production failure that consistently challenge video infrastructure. The first is the re-encoding cascade. A small parameter change in a delivery specification can trigger the need to reprocess an entire content library. What begins as a minor adjustment becomes a multi-week, resource-intensive reprocessing project. Systems that lack predictable output behavior make these cascades worse because operators cannot be certain that reprocessed content will match the original quality profile.

The second category involves audio processing failures. In Arcadian's QA workflows, mislabeled 5.1 surround channels, dual-mono tracks incorrectly flagged as stereo, missing low-frequency effect channels, and unexpected loudness variations account for a significant portion of quality holds. These failures are rarely caught by automated bitstream validation. They require the kind of systematic checking that Arcadian's trained QA specialists perform across every asset.

The third category encompasses adaptive bitrate alignment issues. Misaligned CMAF chunks, variable segment durations across renditions, and discontinuities introduced by server-side ad insertion create playback problems that only manifest on certain devices or network conditions. These are integration failures that exist at the boundary between encoding and delivery, and they multiply in environments serving content across diverse platforms.

KEY INSIGHT

The most expensive failures in video production are not crashes. They are silent quality degradations that pass automated checks but fail the viewer experience. Preventing them requires infrastructure that behaves identically on the thousandth run as it did on the first.

Designing for Simplicity Under Pressure

There is a persistent temptation in infrastructure engineering to optimize for capability. Systems gain features, configuration options multiply, and the operational surface area expands. In lab conditions, this flexibility is an advantage. In production, it is often a liability.

Arcadian's seven-solution service model, spanning transcoding, QA, subtitling, app development, customer support, viewer experience testing, and VPU deployment, gives the company a holistic view of where complexity creates problems. The pattern they observe consistently is that simpler systems outperform more capable ones when conditions deteriorate. A transcoding pipeline with fewer configuration variables produces more consistent results. A deployment that uses standard form factors reduces the number of things that can go wrong during scaling.

This is not an argument against sophistication. It is an argument for placing sophistication in the right layer. VPU silicon handles the computationally complex work of video encoding at the hardware level, which means the software layer above it can remain simpler and more predictable. The operational team interacts with a system that behaves like a well-defined function: input goes in, output comes out, and the relationship between them is stable.

For organizations operating outside controlled data center environments, in constrained network conditions, with heterogeneous hardware, or across geographically distributed processing nodes, this simplicity is not a luxury. It is a prerequisite for reliable operation.

Economics in Production Contexts

The economic case for any infrastructure technology is ultimately proven in production, not in procurement. A lower acquisition cost means nothing if the system requires more operational support. A higher-density solution delivers no value if it introduces reliability problems that consume engineering time.

Arcadian's integration of NETINT VPU technology produced measurable economic results across their operations. The company reports a reduction of more than 50% in video encoding operational expenses, driven by the compound effect of lower power consumption, higher stream density per rack unit, and reduced maintenance overhead. These savings are particularly significant for a global operation with staff and infrastructure spanning Los Angeles, Manila, and Singapore.

But the economic picture extends beyond direct cost reduction. Predictable infrastructure reduces the hidden costs that rarely appear in technology evaluations: the engineering hours spent debugging inconsistent output, the QA cycles consumed by re-verification after unexpected behavior changes, the opportunity cost of capacity that sits idle because operators are uncertain about its reliability. When Arcadian's CEO Joe Waltzer describes the NETINT partnership as a way to deliver exceptional quality at significantly reduced costs, the emphasis belongs on both halves of that equation. The cost reduction is real, but it is enabled by the quality consistency that makes operations more efficient in the first place.

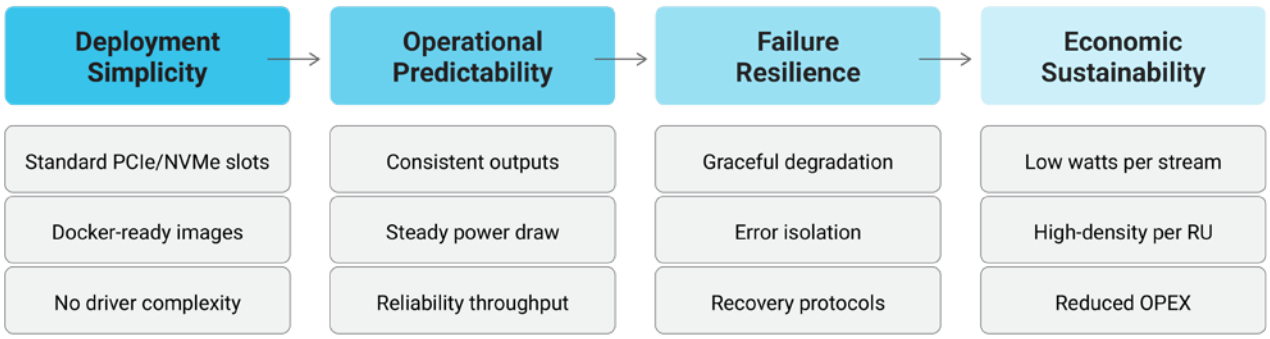


Exhibit 3: Evaluating Production Readiness Across Four Pillars.

Evaluating Production Readiness

For organizations evaluating video infrastructure, the tendency is to start with performance specifications. How many streams per card? What codecs are supported? What is the maximum resolution? These questions matter, but they are insufficient. Production readiness is a broader condition that encompasses deployment simplicity, operational predictability, failure resilience, and economic sustainability.

Deployment simplicity asks whether the solution fits existing infrastructure without specialized accommodation. Standard form factors, established driver ecosystems, and container-ready deployment models all reduce the friction of initial integration and ongoing maintenance. Operational predictability asks whether the system's behavior is consistent enough to build automated workflows around. Failure resilience asks whether the system degrades gracefully when individual components fail and whether errors are isolated rather than cascading. Economic sustainability asks whether the total cost of operation, not just acquisition, supports the business model over the system's expected lifetime.

Arcadian's experience across these four dimensions, earned through years of delivering media services to some of the industry's most demanding clients, suggests that VPU-based infrastructure meets these criteria more completely than the alternatives they have evaluated. The technology is not merely more efficient in isolation. It is more production-ready in aggregate.

The Production Standard

The video infrastructure industry has spent decades optimizing for speed, density, and codec support. These remain important attributes. But the organizations building systems that will operate reliably for years, under conditions they cannot fully predict, are increasingly evaluating a different set of criteria.

They want to know whether the system will behave the same way on day three hundred as it did on day one. They want to know whether adding capacity will be a routine operation or a re-engineering project. They want to know whether the economics will hold up under sustained real-world workloads rather than just during a favorable demo.

Arcadian's partnership with NETINT represents one data point in what appears to be a broader shift in how video infrastructure is evaluated. The companies that have moved VPU technology into production are not doing so because the benchmarks are impressive, though they are. They are doing so because the production results match the benchmark promise, and because the gap between theory and practice, the gap that has plagued infrastructure decisions for years, is narrower with purpose-built silicon than with any alternative they have tested.

That is not a benchmark finding. It is a production finding. In the end, production is the only benchmark that matters.

THIS ARTICLE IS PART OF THE VPU ECOSYSTEM SERIES EXPLORING HOW PURPOSE-BUILT SILICON IS RESHAPING VIDEO INFRASTRUCTURE ACROSS THE TECHNOLOGY STACK. EACH ARTICLE EXAMINES A DIFFERENT ARCHITECTURAL PERSPECTIVE THROUGH THE LENS OF ONE PARTNER WITHIN THE ECOSYSTEM.

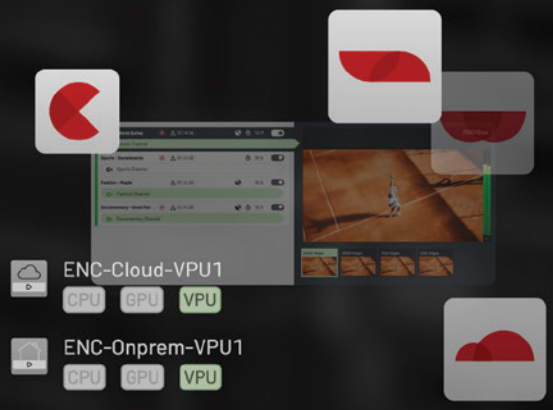
ARCADIAN

Power and Flexibility for Real-World Streaming

Cires21's Streaming Plaform delivers unmatched efficiency and scalability for live streaming. Supporting all major codecs (AV1, H.264, HEVC) and advanced protocols (SRT, NDI, Enhanced RTMP, and more), it ensures seamless, high-quality delivery across any workflow.

Designed for a truly hybrid environment, the platform brings together all essential streaming capabilities into a unified solution, operating both on-premises and in the cloud to adapt to any requirement. With Live Control at the core, it maximizes performance and efficiency at every stage of the process.

It has been measured against real workloads for real operators and broadcasters like **Real Madrid**, **Dorna Sports (MotoGP)**, **Mediaset**, **RTVE** and **Eurovision Services** for more than 18 years.



18+
Years of experience

Up to 16
SDI on a single rack unit

5×
More efficient than GPU when using NETINT's VPUs

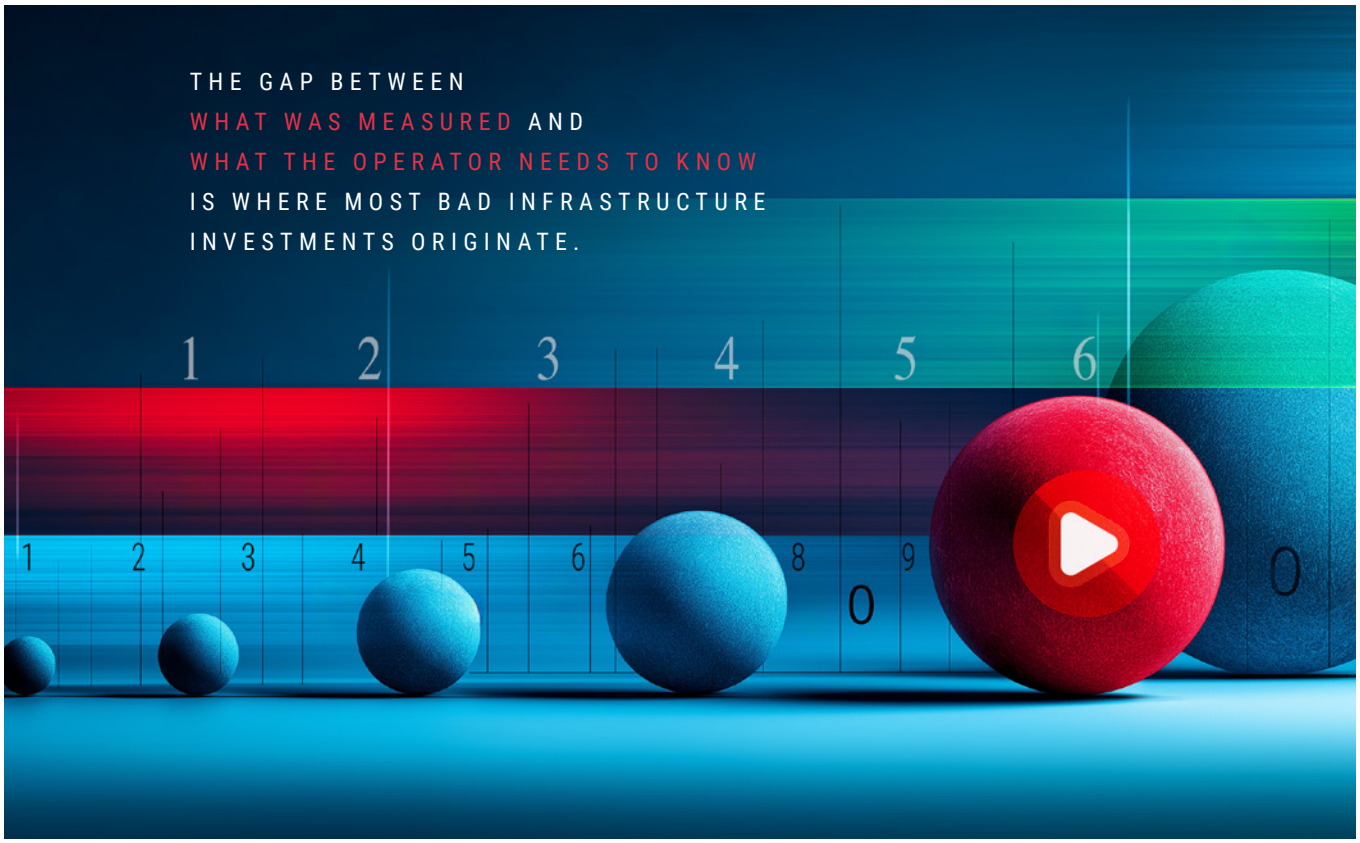
MEDIA PROCESSING

Measuring What Matters

How Cires21 is Turning Video Benchmarking into a Discipline Worth Trusting.

There is a quiet crisis in video infrastructure, and it has nothing to do with codecs, bitrates, or buffering ratios. It is a crisis of measurement. As streaming operators weigh decisions that carry multi-year capital implications, they increasingly rely on performance benchmarks that were never designed to bear that weight. The numbers look precise. The methodologies are anything but.

This is a problem that Cires21, a Madrid-based video technology company with more than fifteen years of experience in broadcast and streaming, understands intimately. Founded in 2008, Cires21 has built encoding and delivery infrastructure for some of Europe's most demanding video environments, serving organizations such as RTVE (Spain's national broadcaster), Dorna Sports, Telefonica, Mediaset, Real Madrid, Atresmedia, and Eurovision Services. When the company began evaluating hardware-accelerated encoding alternatives, it realized that the industry's standard benchmarking practices were failing to capture what operators actually need to know.



Why Measuring Video Is Hard

Video encoding is one of those domains where complexity hides behind familiar metrics. A bitrate number seems straightforward. A VMAF score looks like it settles the quality question. But both are abstractions layered on top of a system with dozens of interacting variables, and each abstraction strips away context that matters.

Consider what happens during a live encoding session. The encoder must make frame-by-frame decisions under real-time constraints, adapting to scene changes, motion intensity, and content complexity. It must do this while managing thermal limits, power budgets, and latency requirements that shift depending on the delivery target. A benchmark that captures one of these dimensions while ignoring the others is misleading.

Cires21's C21 Live Encoder supports codecs range from H.264, HEVC, AV1, VP9, and MPEG-2 across protocols including HLS, DASH, CMAF, SRT, and WebRTC, has to perform under all of these conditions simultaneously. That breadth of real-world deployment experience is precisely what led the company to question whether the industry's measurement conventions were keeping pace with the technology they were meant to evaluate.

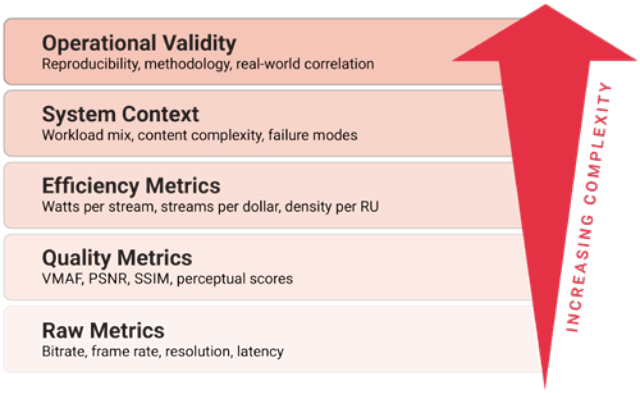


Exhibit 1:
Each layer of the measurement stack introduces variables that simpler benchmarks tend to ignore.

The Limits of Headline Metrics

The most common benchmarking pitfall is what might be called the "single-number fallacy." An encoder achieves a VMAF score of 95 at a given bitrate. Another achieves 93. The first must be better. But this reasoning collapses under scrutiny when you consider what was left unmeasured.

Was the content representative of actual production workloads, or was it selected to flatter one architecture? Were the encoders configured with equivalent optimization effort, or did one receive weeks of tuning while the other ran with defaults? Was power consumption tracked per stream, or was the total system draw buried in a footnote? Did the quality metric capture tail performance (the worst 5% of frames that viewers actually notice), or did a generous average mask inconsistency?

This is not a theoretical concern. Cires21's engineering team, led by CTO David Gonzalez, has documented specific cases where standard industry benchmarks produced rankings that reversed under real-world operating conditions. An encoder that won a controlled lab comparison consumed three times the power per stream in continuous 24/7 operation. A codec configuration that excelled on test clips produced unacceptable quality variance on broadcast content with frequent scene changes.

Efficiency as a System Property

One of the most important conceptual shifts in Cires21's approach is treating efficiency not as a feature of individual components but as a property of the entire system. A chip that encodes streams at low wattage is only efficient if it maintains quality, handles the required density, and operates within the thermal envelope of a production rack. Efficiency is the relationship between all of these variables, not a number that lives on a datasheet.

This perspective becomes especially important when evaluating purpose-built silicon like VPUs (Video Processing Units) against general-purpose GPUs. The two architectures make fundamentally different engineering trade-offs, and comparing them requires a framework that accounts for those differences rather than ignoring them.

Cires21 confronted this directly when it began integrating NETINT Quadra T1U VPUs into the C21 Live Encoder. The encoder had evolved through three generations of acceleration: a CPU era where software encoding dominated, a GPU era that brought density improvements at higher power costs, and beginning in 2024, a VPU era that promised a different set of trade-offs entirely. Each transition demanded new measurement approaches because the previous generation's benchmarks could not fairly evaluate the new architecture.

KEY INSIGHT

A benchmark that answers only the question it was designed to ask may still mislead the person using it to make a different decision entirely. The gap between "what was measured" and "what the operator needs to know" is where most bad infrastructure investments originate.

Building a Better Benchmark

Rather than accepting industry convention, Cires21 developed a structured methodology for designing what it calls "fair benchmarks for heterogeneous video compute." The framework centers on four principles that sound obvious in theory but are rarely followed in practice.

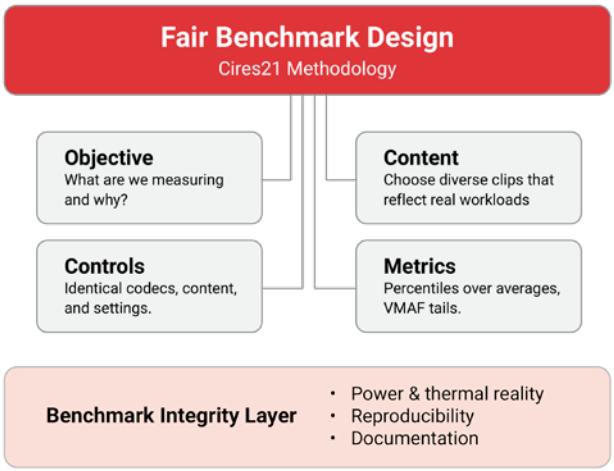


Exhibit 2: Cires21's four-pillar methodology for designing benchmarks that hold up under operational scrutiny.

First, objective definition. Before any test runs, the team establishes precisely what question the benchmark is designed to answer. "Which encoder is better?" Is not a valid objective. "Which architecture delivers the lowest cost per stream at broadcast quality under 24/7 operation?" Is. The specificity forces everyone involved to confront the trade-offs before the numbers are generated rather than after.

Second, normalization and controls. Every comparison must use equivalent codec settings, equivalent optimization effort, and equivalent operational conditions. When Cires21 benchmarked the NETINT Quadra T1U against the NVIDIA RTX 4000 Ada, both platforms received the same tuning methodology applied by engineers with comparable expertise on each architecture. The VisualON Optimizer was integrated to ensure perceptual quality optimization was consistent across platforms.

Third, content selection. Test clips must represent the diversity of actual production workloads: high-motion sports content, talking-head interviews, graphics-heavy broadcasts, and low-light footage. Using only friendly content (static scenes, low complexity) flatters every encoder and tells the operator nothing about behavior under stress.

Fourth, metric selection. Cires21 insists on percentile-based quality reporting rather than averages. A mean VMAF of 95 can hide a long tail where 5% of frames fall below 85, which is exactly the kind of artifact that viewers notice and complain about. The 5th percentile VMAF score is often a more honest indicator of viewer experience than the mean.

What the Numbers Revealed

When Cires21 applied this methodology to its VPU evaluation, the results were instructive. The joint white paper with NETINT, titled "More Streams, Less Power: Energy Trade-offs," benchmarked the Quadra T1U against the NVIDIA RTX 4000 Ada across H.264, HEVC, and AV1 encoding workloads.

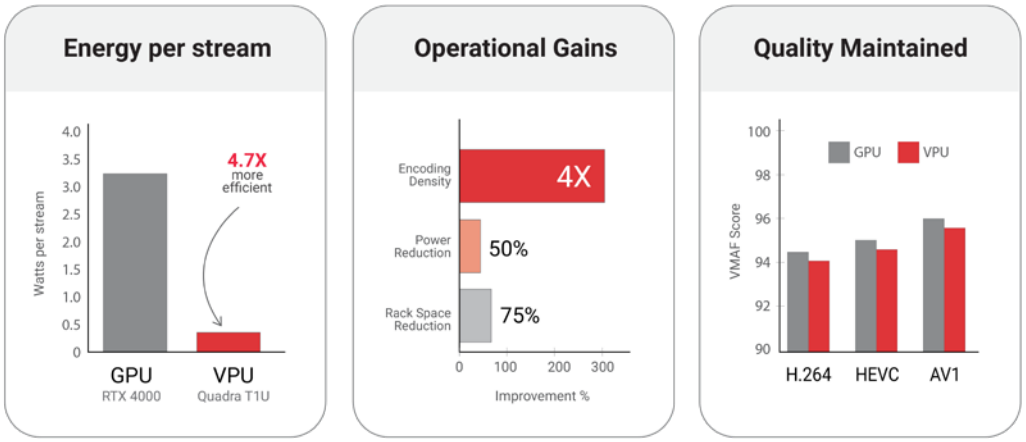


Exhibit 3: Benchmark results from the Cires21-NETINT white paper across energy, operations, and quality dimensions.

The headline finding was striking: VPUs delivered 4.7 times greater energy efficiency, consuming between 0.4 and 0.7 watts per stream compared to significantly higher consumption on the GPU platform. But the methodology's value showed in the supporting data. Quality was not sacrificed for efficiency. VMAF scores remained comparable across both architectures and all three codecs. The benchmarks included percentile distributions, not just averages, confirming that tail quality was maintained.

At the operational level, the numbers translated into concrete deployment advantages: a 75% reduction in rack space, a 50% decrease in power consumption, and a fourfold increase in encoding density. These are the metrics that infrastructure planners actually use to model total cost of ownership, and Cires21 published them alongside the raw performance data so operators could map the results to their own environments.

From Benchmarks to Deployment Decisions

The real test of any measurement framework is whether it survives contact with production. Cires21's experience deploying VPU-based encoding across its customer base suggests that the benchmarks held. The C21 Live Encoder now operates in environments ranging from national broadcasting (RTVE, TV3, EITB) to sports streaming (Dorna Sports, Real Madrid) to telecommunications infrastructure (Telefonica, Cellnex Telecom), and the transition from GPU to VPU acceleration has proceeded without the quality regressions that poorly characterized technology transitions typically produce.

This matters because the video industry is entering a period of significant infrastructure renewal. The migration to AV1, the growth of live-to-VOD workflows through tools like Cires21's C21 Live Editor, the expansion of AI-driven content processing through platforms like MediaCopilot, and the continued push toward lower latency via WebRTC (WHIP/WHEP) all place increasing demands on encoding infrastructure. Operators making five-year capital investments need benchmarks they can trust, not benchmarks that were designed to close a sale.

THE BIGGER PICTURE

Efficiency benchmarks only earn trust when they are accompanied by the full measurement context: content diversity, quality percentiles, power methodology, and documentation sufficient for an independent team to reproduce the results. Cires21's published methodology meets this standard.

Common Pitfalls and How to Avoid Them

Cires21's methodology documentation identifies several measurement traps that are worth highlighting because they recur across the industry.

The optimization parity problem: When comparing two architectures, it is common for the incumbent to receive extensive tuning while the challenger runs with minimal optimization. This produces results that reflect the effort invested in tuning, not inherent capability of the hardware. Cires21's protocol requires that each platform receive equivalent engineering attention, documented and auditable.

The friendly-content trap: Benchmarks that use only low-complexity test sequences overstate the performance of every encoder and obscure the differences that matter most. Production content is unpredictable, and measurement must reflect that unpredictability. Cires21 mandates a content mix that includes challenging material like sports with fast camera motion, concerts with strobe lighting, and news tickers with sharp text overlays.

The averages-vs-percentiles blind spot: A mean VMAF of 95 across a session can coexist with individual frames scoring below 80. Viewers do not experience averages. They experience the worst frames. Any quality claim that reports only mean scores without 5th or 1st percentile data is incomplete, and Cires21's framework treats it as such.

The power-measurement gap: Many benchmarks omit power data entirely or report it at the wall level without isolating the encoding workload. Cires21's methodology requires per-stream power measurement under sustained load, because power consumption often increases non-linearly as density rises and thermal management becomes a factor.

Measurement as a Competitive Advantage

There is an emerging argument that the organizations with the best measurement practices will make the best infrastructure decisions, and that this advantage compounds over time. Operators who invest in rigorous evaluation avoid expensive technology bets that underperform in production. They build vendor relationships grounded in transparency rather than marketing claims. They create internal knowledge that makes each subsequent decision faster and better informed.

Cires21's trajectory illustrates this pattern. By publishing its benchmarking methodology rather than treating it as proprietary, the company has positioned itself as a trusted evaluation partner, not just a technology vendor. The video industry does not lack for performance data. What it lacks is performance data that operators can trust, contextualize, and reproduce. Cires21's contribution is not simply another set of benchmark numbers. It is a framework for deciding which numbers deserve to be believed so operators can get the technical results they need.

THIS ARTICLE IS PART OF THE VPU ECOSYSTEM SERIES EXPLORING HOW PURPOSE-BUILT SILICON IS RESHAPING VIDEO INFRASTRUCTURE ACROSS THE TECHNOLOGY STACK. EACH ARTICLE EXAMINES A DIFFERENT ARCHITECTURAL PERSPECTIVE THROUGH THE LENS OF ONE PARTNER WITHIN THE ECOSYSTEM.

DELL Technologies

DATA CENTER HARDWARE

From Silicon to Rack: What It Takes To Make Video Processing Deployable At Data Center Scale

Efficient Silicon Is Necessary But Not Sufficient. Why Deployable Video Infrastructure Demands a New Design Philosophy.

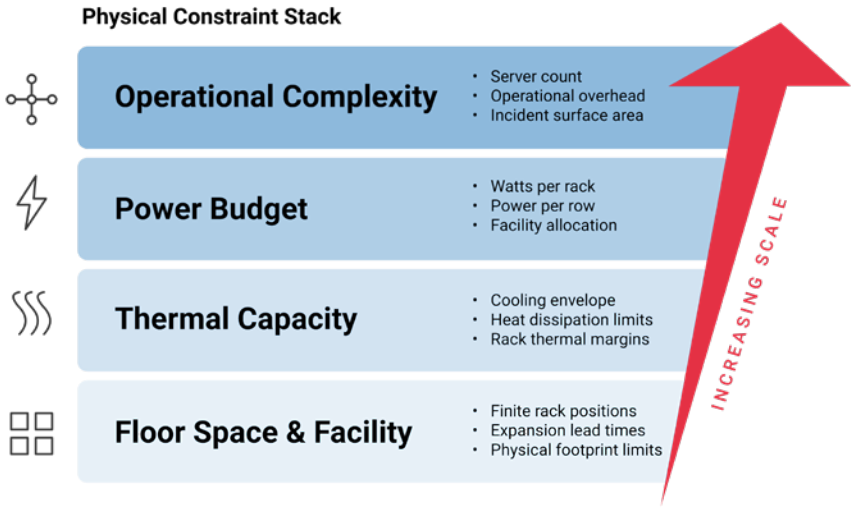
The Abstraction Breaks Down

At a modest scale, video infrastructure decisions can remain largely abstract. Engineering teams optimize for throughput or quality and assume the underlying systems will accommodate whatever they build. When concurrency is limited and workloads are intermittent, inefficiency shows up primarily as a higher bill.

At production scale, that assumption stops holding. Video workloads are sustained, power-intensive, and concurrency-driven. As deployments grow, teams hit constraints that no amount of software optimization

Exhibit 1: At scale, these limits compound and become non-negotiable.

can solve: power budgets per rack, thermal dissipation limits, floor space availability, and the sheer operational costs of maintaining large server fleets. In these environments, inefficiency does not simply raise costs. It determines whether a given workload can physically exist within real data center constraints. The question shifts from how fast, to how deployable?



Why General-Purpose Servers Struggle

Historically, video processing has relied on general-purpose compute, with GPUs added when acceleration was needed. That model provides flexibility, but it introduces challenges when applied to dense, sustained video workloads.

CPU- and GPU-based encoding often exhibits fluctuating power draw depending on workload mix and tuning. High-performance accelerators can concentrate heat in ways that complicate airflow and cooling design. Fewer streams per server means more rack space, more cabling, and more operational overhead. And capacity planning becomes probabilistic rather than deterministic, forcing infrastructure teams to over provision power, cooling, and space to accommodate worst-case scenarios.

The result is a planning model built on estimates and margins rather than known, repeatable behavior. Every variable that fluctuates adds a layer of uncertainty to facility design, procurement, and operations.

KEY INSIGHT

At production scale, the distinction between a variable workload and a predictable one is not a matter of preference. It is the difference between over provisioning by 30-40% and sizing to actual demand.

VPUs Change the Planning Model

Within the VPU Ecosystem, Video Processing Units shift the planning model from theoretical performance to predictable physical behavior. From an infrastructure perspective, three changes matter most.

Deterministic power profiles. Dedicated video silicon delivers stable, bounded power consumption per stream. This simplifies power budgeting at the server and rack level. Teams can plan around known energy draw and compute capacity rather than peak estimates.

Higher density per footprint. More streams per server means fewer systems are needed to deliver a given level of video encoding capability. Rack count drops. Cabling simplifies. Operational overhead shrinks.

Thermal predictability. Consistent thermal output improves cooling efficiency and reduces the need for conservative design margins. Airflow planning becomes straightforward rather than defensive.

Together, these characteristics let video infrastructure teams plan capacity based on concrete physical constraints rather than peak-performance estimates. With VPUs, the planning model shifts from probabilistic to deterministic.

Planning model comparison

	LEGACY MODEL	NETINT MODEL
	General-purpose	VPU-accelerated
Power Profile	Variable, workload-dependent	Deterministic, bounded
Stream Density	Lower streams per server	Higher streams per server
Thermal Behavior	Concentrated, unpredictabl	Consistent, predictable
Capacity Planning	Probabilistic, peak-based	Deterministic, linear
Provisioning	Overprovision for worst case	Right-size to actual demand

Exhibit 2: How VPU-accelerated infrastructure changes planning assumptions across five key dimensions.

From Silicon to System: Where Server Design Earns Its Keep

Efficient silicon alone is not sufficient. To be deployable at scale, acceleration must be integrated into systems designed for reliability, serviceability, and lifecycle management.

Server design for sustained video workloads involves considerations that go well beyond raw performance: chassis airflow and component placement, power supply sizing and redundancy, PCIe layout and expansion flexibility, firmware stability and platform validation, and long-term serviceability across multi-year deployments.

This is the kind of engineering that Dell Technologies has refined across decades of enterprise server design. The PowerEdge platform, for example, offers PCIe Gen5 expansion across a range of form factors. The R760 supports up to eight PCIe slots with a mix of Gen4 and Gen5 risers, giving infrastructure teams the flexibility to pair NVMe VPU accelerators alongside networking and storage cards without slot contention. For denser configurations, the XE series delivers higher accelerator counts in validated thermal envelopes. The underlying principle is the same: provide enough expansion headroom so that adding purpose-built acceleration does not require rearchitecting the server.

Thermal management is equally critical. Dell's Multi-Vector Cooling (MVC) framework combines hardware innovations with AI-driven adaptive fan control. The latest generation uses a fuzzy-logic closed-loop controller that adjusts fan speed based on real-time thermal sensor input and power telemetry, optimizing for both cooling performance and acoustic efficiency. For video workloads, where sustained thermal output is the norm rather than the exception, this kind of adaptive control keeps components within safe operating ranges without the energy waste of running fans at full speed continuously.

In this context, VPUs are not treated as experimental accelerators bolted onto general-purpose hardware. They are production components within validated server platforms, designed to operate reliably under sustained load for years at a time. This distinction matters. Experimental hardware often works in the lab but creates problems at scale: inconsistent firmware behavior, unvalidated thermal profiles, unclear support paths. Production-grade integration eliminates these risks before they reach the data center floor.

Density Changes Operational Economics

When stream density increases, the operational math changes at every level. Fewer servers mean fewer rack positions. Fewer racks mean less power, less cooling, and less floor space. Fewer systems to manage means fewer failure points, less cabling, and simpler monitoring.

This is not a marginal improvement. When you move from dozens of streams per server to hundreds, you are not optimizing an existing model. You are changing the category of infrastructure required. Deployments that once filled rows of racks can fit in a fraction of the space, with proportional reductions in operational complexity.

Dell's iDRAC management platform amplifies this advantage. Each PowerEdge server includes an integrated remote access controller capable of streaming real-time telemetry, covering power consumption, thermal sensors, fan speeds, and component health, via the Redfish API. For video deployments, where dozens of servers may run identical sustained workloads, this telemetry makes it straightforward to monitor fleet health from a single pane. When a VPU's thermal signature shifts or a power supply nears its rated capacity, operations teams can catch the issue before it becomes a service disruption.

For organizations operating multiple facilities or expanding into new regions, this density advantage compounds. Each new deployment starts from a smaller, simpler, more predictable baseline.

Planning for Lifecycle, Not Just Launch

Video infrastructure is rarely static. Platforms evolve, codecs change, and demand grows over time. Deployable systems must account for multi-year support horizons, incremental capacity expansion, hardware refresh cycles, and operational continuity during upgrades. Predictable, efficient workloads reduce the friction at every stage of this lifecycle. When capacity scales linearly and power behavior is stable, infrastructure planning becomes iterative rather than disruptive. A hardware refresh is an upgrade, not a redesign. A capacity expansion is a known quantity, not a new engineering project.

Dell's OpenManage Enterprise suite is designed to automate precisely these lifecycle tasks across large fleets. It can manage up to 8,000 PowerEdge devices from a single console, handling firmware updates, configuration compliance, and hardware health monitoring through policy-based automation.

For video operators who may be deploying dozens or hundreds of VPU-equipped servers, the ability to push firmware updates, validate configurations, and track hardware health across an entire fleet without touching individual machines is a material reduction in operational burden. Combined with Ansible integration for infrastructure-as-code workflows, the management layer scales alongside the hardware.

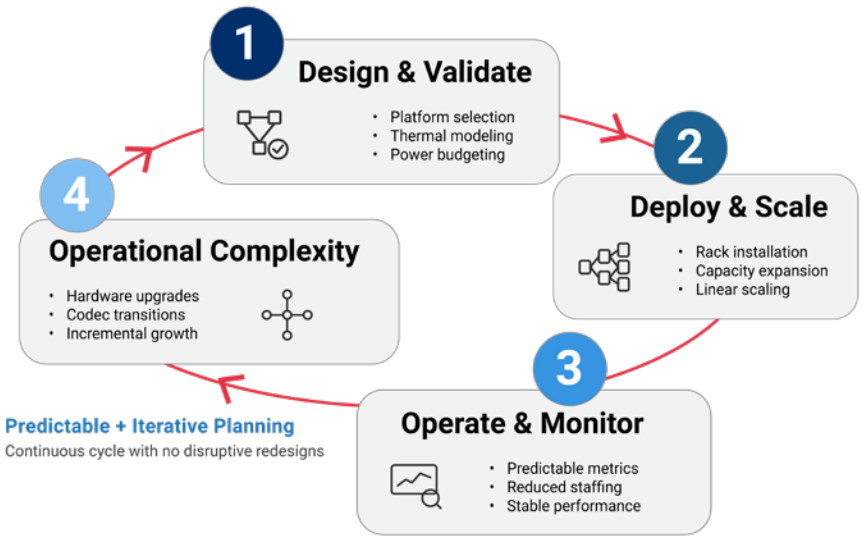


Exhibit 3: The Infrastructure Lifecycle. Predictable workloads make each phase iterative rather than disruptive.

THE BIGGER PICTURE

When servers are designed around predictable, efficient video processing, higher-level architectural decisions become easier. Compute can be placed closer to users when latency matters. Capacity can scale without rearchitecting facilities. Operational teams can support video growth without proportional increases in complexity.

Infrastructure Is the Foundation

Efficiency in video processing is often discussed in terms of cost or performance. At infrastructure scale, it becomes a question of feasibility. Can this workload physically exist in this facility? Can it be supported over time? Can it scale without rebuilding the environment around it?

From silicon to rack, deployable video infrastructure depends on predictable behavior, density, and supportable system design. VPUs change what is possible at the hardware level. Server platforms translate that efficiency into systems that can be deployed, operated, and evolved over time. When the silicon is purpose-built, like the VPU is for video, and the server platform is engineered for sustained, real-world operation, what once required rows of general-purpose machines can fit in a fraction of the space, with a fraction of the operational overhead. This is why the world's largest video platforms and social networks all use custom silicon to encode and process video.

Within the VPU Ecosystem, infrastructure is not an afterthought. It is the foundation that makes efficiency real.

THIS ARTICLE IS PART OF THE VPU ECOSYSTEM SERIES EXPLORING HOW PURPOSE-BUILT SILICON IS RESHAPING VIDEO INFRASTRUCTURE ACROSS THE TECHNOLOGY STACK. EACH ARTICLE EXAMINES A DIFFERENT ARCHITECTURAL PERSPECTIVE THROUGH THE LENS OF ONE PARTNER WITHIN THE ECOSYSTEM.

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:

EXECUTIVE SUMMARY

6 Findings Defining 2026 Video Encoding Market

Cost and quality are no longer trade offs

This year's data reveals an industry navigating simultaneous pressures: 46% of respondents were VP-level or above, validating strategic level findings, while 35% were engineers and architects grounding operational insights. 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.

ONE

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.

TWO

GPU dominance creates 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.

THREE

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.

FOUR

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.

FIVE

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.

SIX

Executives and engineers align on cost, yet 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.

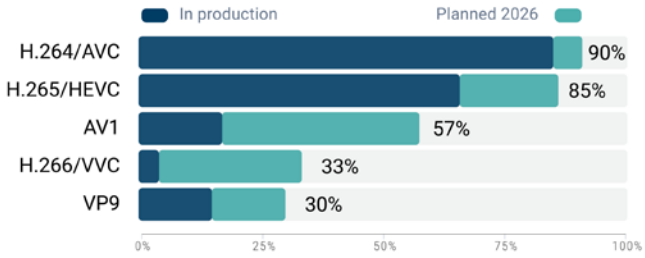
Position solutions against revenue growth and unit economics—the dominant executive KPIs.

EXECUTIVE SUMMARY

Technology Snapshot & Strategic Priorities for 2026

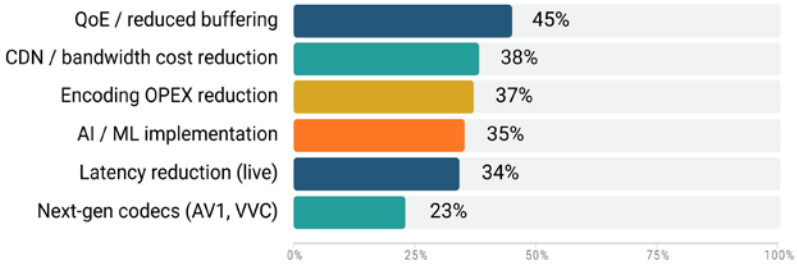
Codec Adoption Trajectory

Production deployment vs. 2026 combined reach.



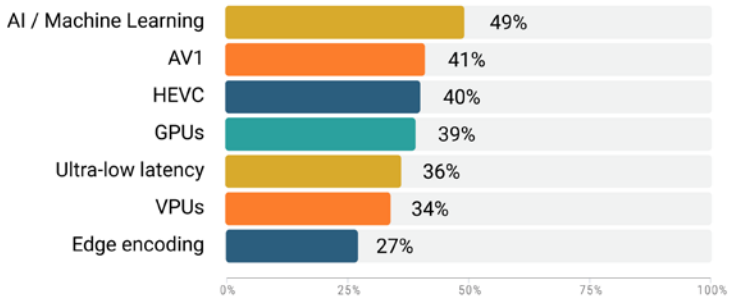
Strategic Priorities

Top technology initiatives.



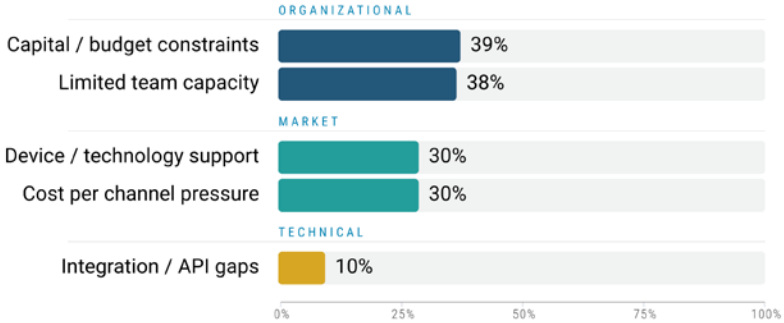
Expansion Plans

Top technology expansion area (% planning).



Roadmap Blockers

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



EXECUTIVE SUMMARY

4 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 vendor approaches.

37% OF SAMPLE

Small Team Operators

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

26% OF SAMPLE

Specialized / Evaluators

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

24% OF SAMPLE

Infrastructure Engineers

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

13% OF SAMPLE

Enterprise Hyperscalers

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

INDUSTRY REPORT 2026

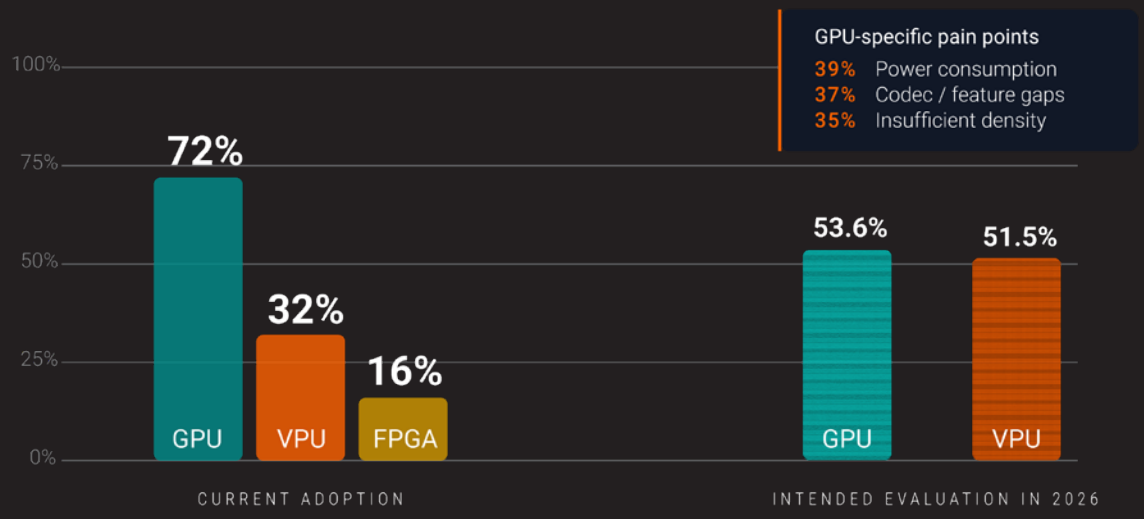
10 Insights That Matter

A strategic synopsis tailored for the VPU Ecosystem, condensed from the 47-page 2026 State of Video Encoding Industry Report *Q4 2025*.

10 INSIGHTS THAT MATTER

01 The GPU Monoculture Is Cracking

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 evaluation gap

41% Deploy multiple hardware types

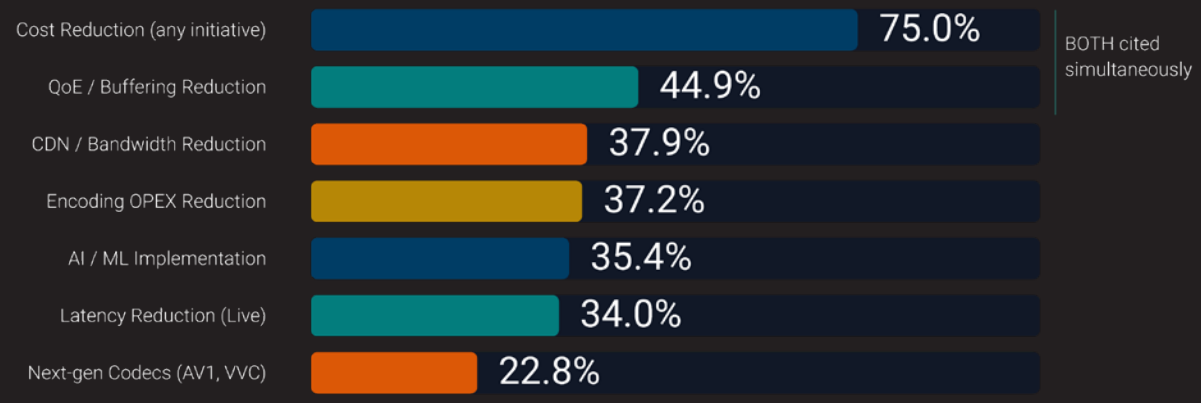
39% GPU users: power pain

72% GPU current adoption

10 INSIGHTS THAT MATTER

02 The Dual Mandate Has Replaced the Tradeoff

The old market logic — sacrifice quality for cost, or spend for quality — 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.



2.31
Avg initiatives per org

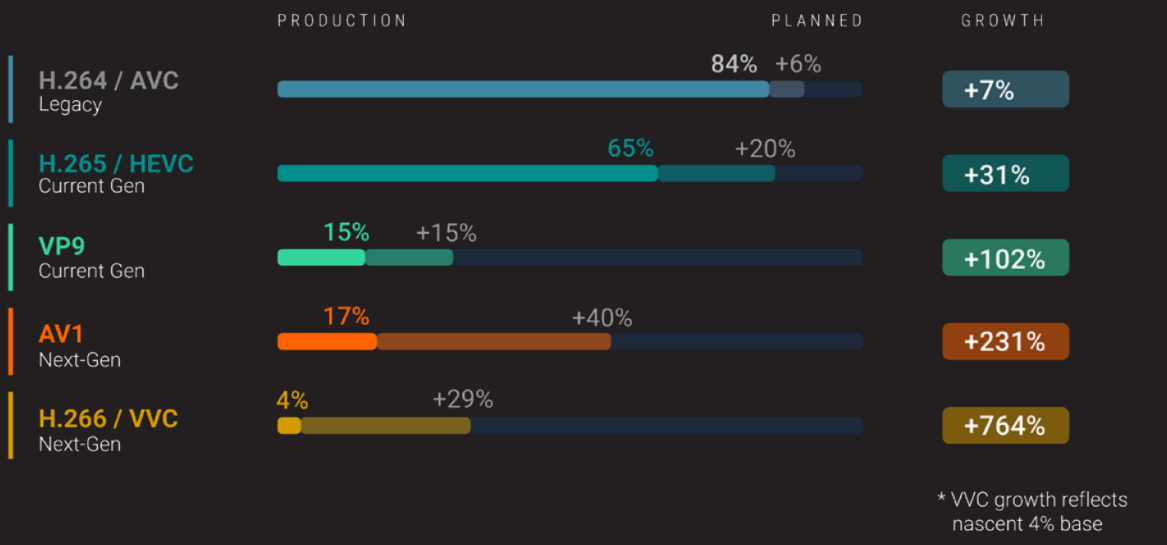
42%
Select all 3 priorities

75%
Cite ≥1 cost initiative

10 INSIGHTS THAT MATTER

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

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. Combined reach hits 57% by year-end. 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.



* VVC growth reflects nascent 4% base

57%
AV1 combined reach EOY 2026

27x
VP9 orgs more likely with AV1

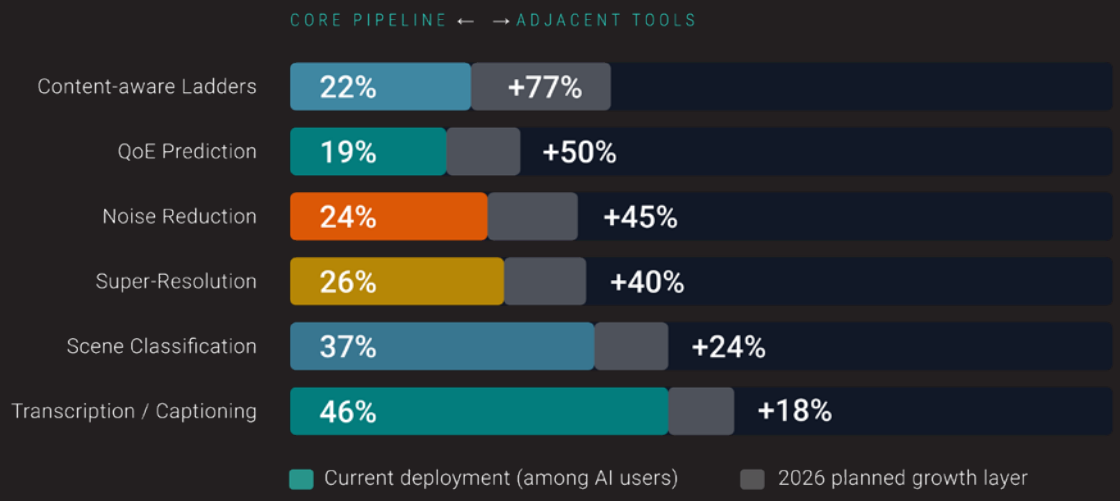
+231%
AV1 planned growth rate

54%
Hardware decode still a barrier

10 INSIGHTS THAT MATTER

04 AI Has Crossed from Experiment to Infrastructure

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 double 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.



43.8%
Named AI as future force (unprompted)

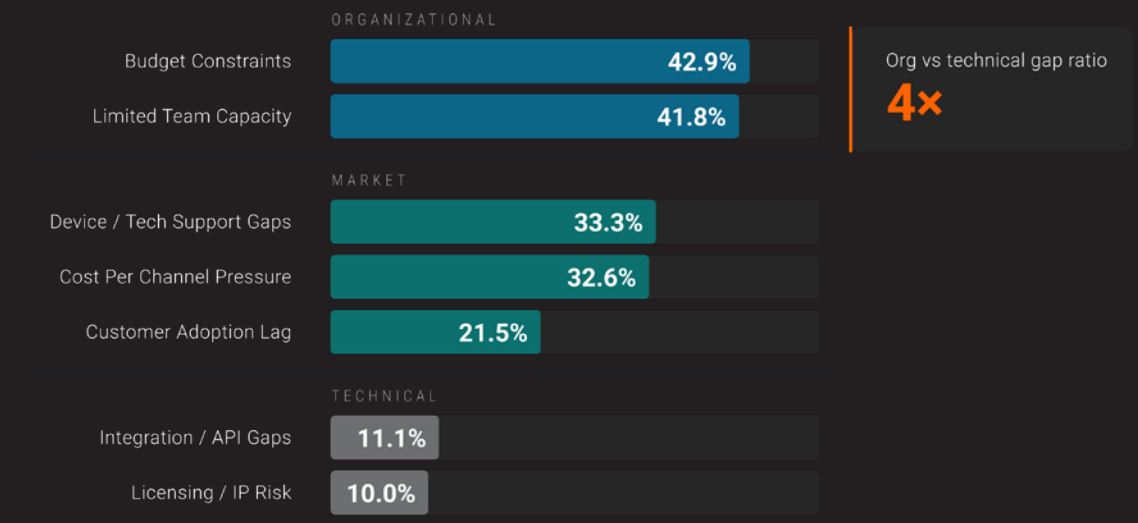
53%
Plan AI expansion in 2026

60%
Already deploy AI / ML

10 INSIGHTS THAT MATTER

05 The Binding Constraint Is Organizational, Not Technical

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 not a feature. It is the category.



77%
Cite ≥1 organizational barrier

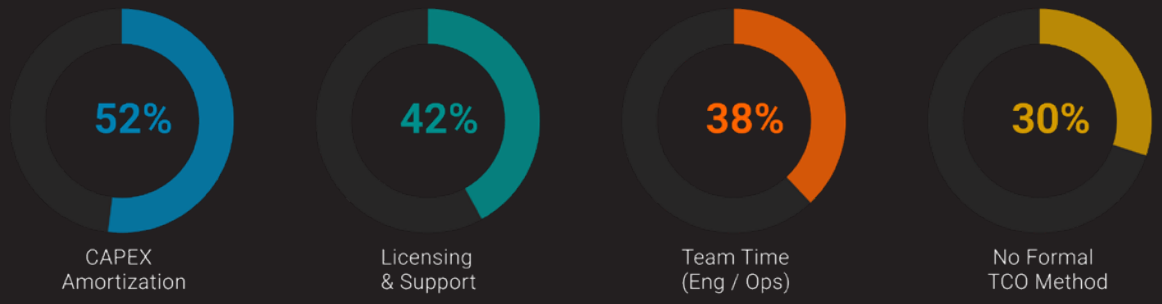
4x
Organizational vs. technical ratio

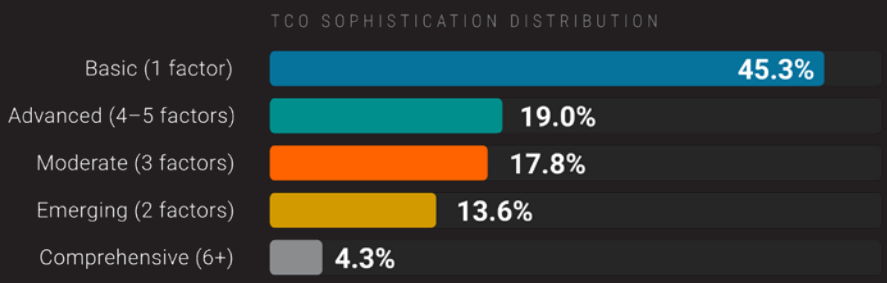
49%
Teams of 5 or fewer people

10 INSIGHTS THAT MATTER

06 TCO: The Industry's Biggest Vulnerability

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 way to measure wins their trust before any benchmark conversation begins.





30% No TCO methodology

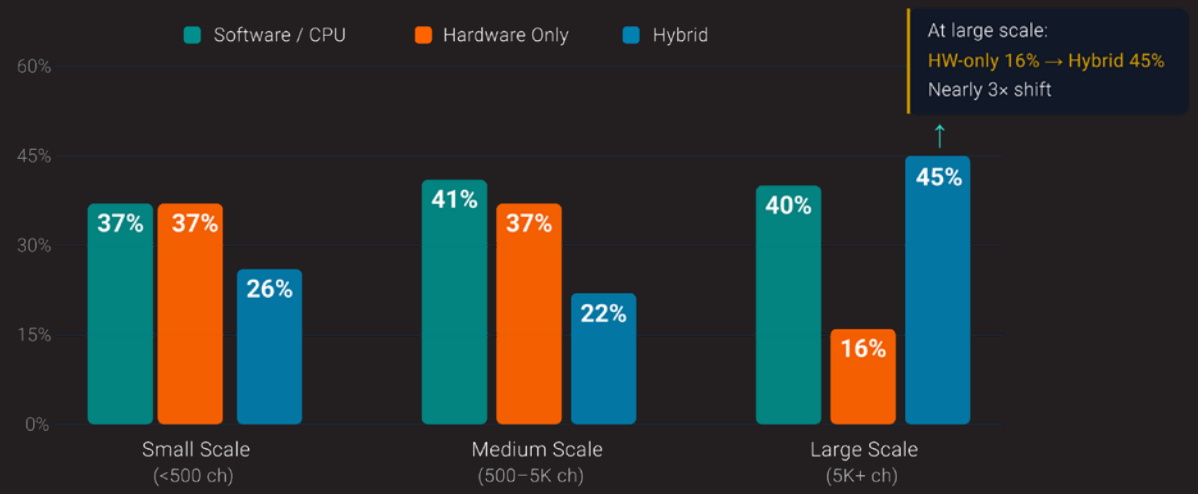
27% Track zero unit economics

37.4% Methodical evaluators hit hardest

10 INSIGHTS THAT MATTER

07 At Scale, Hybrid Wins. Hardware-Only Loses.

The scale-to-method relationship is unambiguous. 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.



16%
Hardware only at large scale

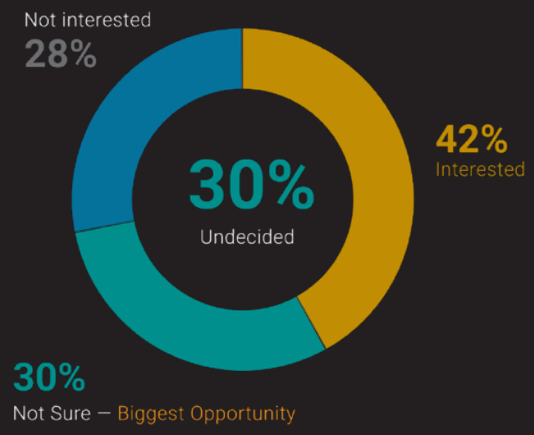
45%
Hybrid at large scale

56%
Super operators use hybrid

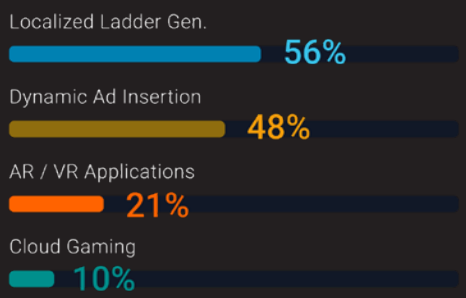
10 INSIGHTS THAT MATTER

08 Edge Is a $0 Market Waiting for a Reason

42% of respondents express interest in edge encoding. 30% say they are "not sure." That uncertainty band, nearly one in three professionals, is not skepticism. It is 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



EDGE INTEREST BY OPERATIONAL SCALE

250+ live channels	50%
1–10 live channels	35%

42%
Interested in edge encoding

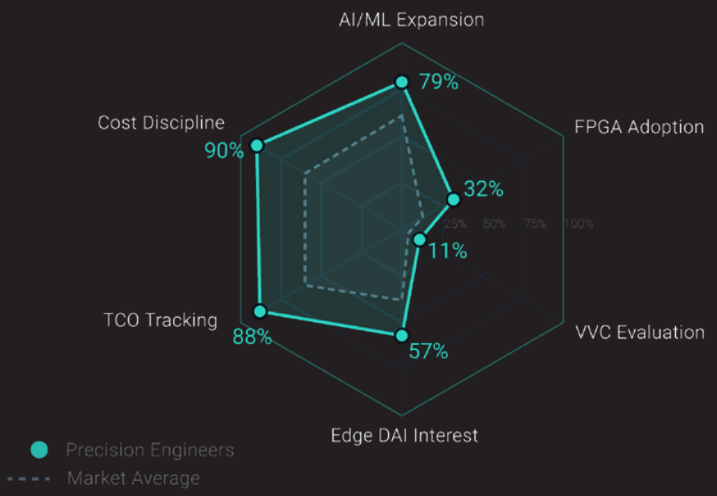
30%
Uncertain. Educational opportunity.

56%
Want localized ladder generation

10 INSIGHTS THAT MATTER

09 Precision Engineers Are Small in Number, Large in Influence

At 12.9% of the market, Precision Engineers — live sports, betting, mission-critical broadcast — punch well above their segment share. They lead on AI/ML expansion intent (78.8%), FPGA adoption, VVC evaluation, edge DAI interest, and TCO sophistication. They are the lighthouse customers whose infrastructure decisions the rest of the market watches and follows. Winning here does not just close deals — it shapes the category narrative.



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%	78.8%	0%
Market share	AI / ML expansion intent	Outsourced. All build or hybrid.

10 INSIGHTS THAT MATTER

10 Four Distinct Buying Personalities

Statistical clustering reveals four archetypes: Methodical Evaluators (37.4%) require lab hardware and documented proof. Scale Operators (25.5%) need turnkey integration and executive ROI language. Agile Innovators (24.1%) will try cloud-first and move fast. Precision Engineers (12.9%) demand latency specs, flexibility guarantees, and measurable cost models. Cloud PoC acceptance ranges from 10.7% to 41.3% depending on archetype — a 4x gap. One evaluation path, one message, one motion captures one quarter of the market.

Methodical Evaluators 37.4%

60%	Lab PoC required
11%	Cloud PoC accepted
34.6%	Untracked unit economics
Long	Sales cycle

Show proof. Not features.

Scale Operators 25.5%

53%	Lab PoC required
60%	AI/ML expansion intent
28.7%	Highest VPU interest
Med–Long	Sales cycle

Connect tech to exec KPIs.

Agile Innovators 24.1%

41%	Cloud PoC accepted
74%	AI/ML expansion intent
37.7%	SaaS evaluation interest
Short–Med	Sales cycle

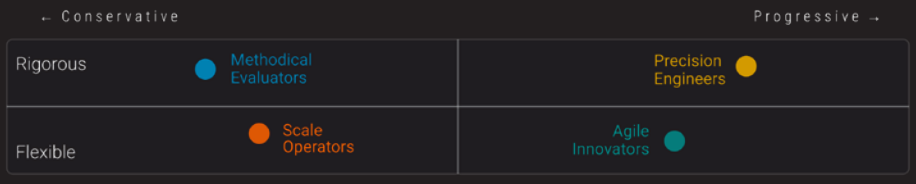
Cloud trial → fast convert.

Precision Engineers 12.9%

79%	AI/ML expansion intent
2.6	TCO factors tracked (highest)
57%	Edge DAI interest
Medium	Sales cycle

Lighthouse customers. Win the narrative.

EVALUATION POSTURE MAP



37.4% Methodical Evaluators, largest segment

4x Cloud PoC gap (10.7% - 41.3%)

12.9% Precision Engineers, highest influence

10 Insights That Matter

The through-line across all ten:

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.

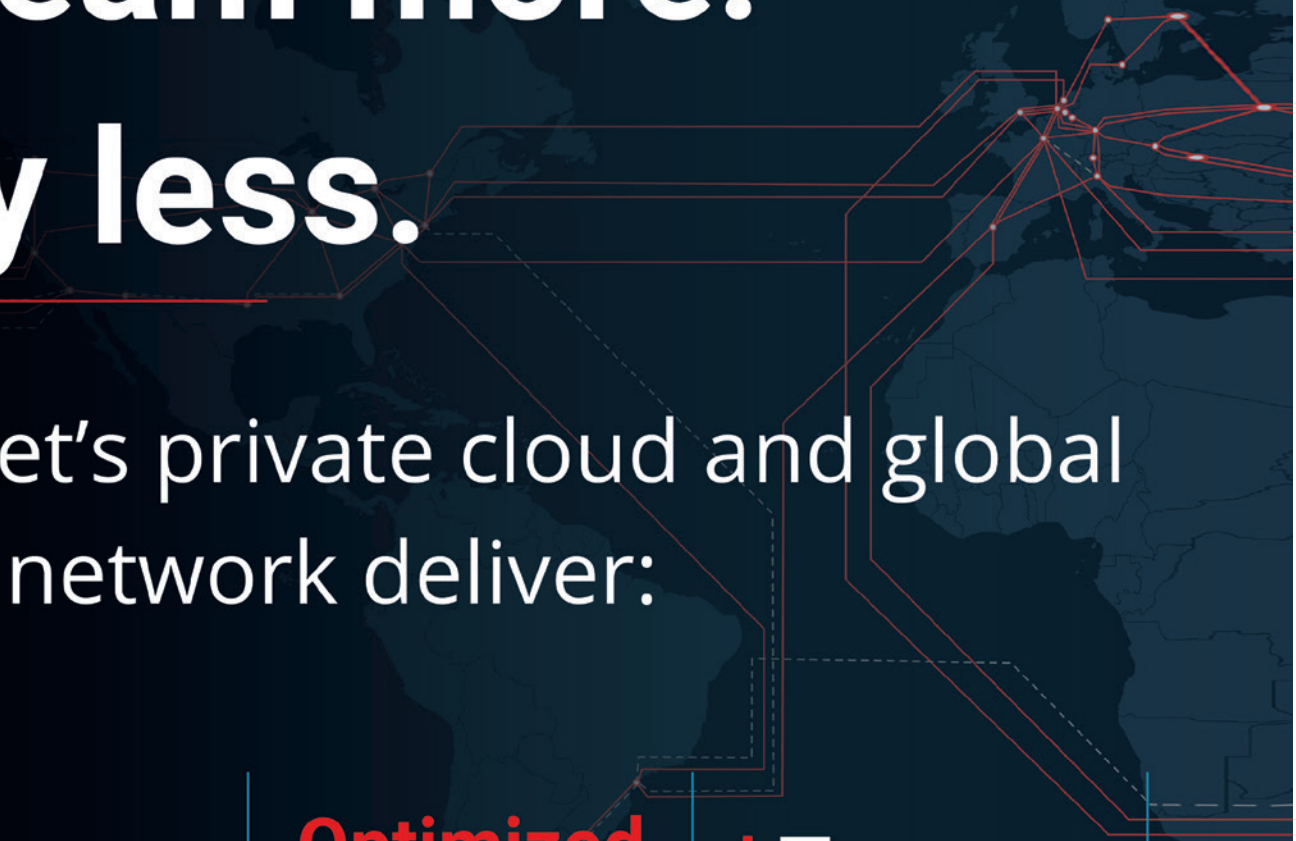
That is the category NETINT is building. These 10 data points are the demand proof.

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

Link to unabridged report:

i3D.net

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

Visit us at NAB - Test a server for 30 days

Talk to our team to get started.

REAL-TIME INFRASTRUCTURE

Where Video Compute Belongs: A Latency-First Perspective

How i3D.net and NETINT VPUs Are Making Distributed Transcoding Viable For Latency-Sensitive Workloads.

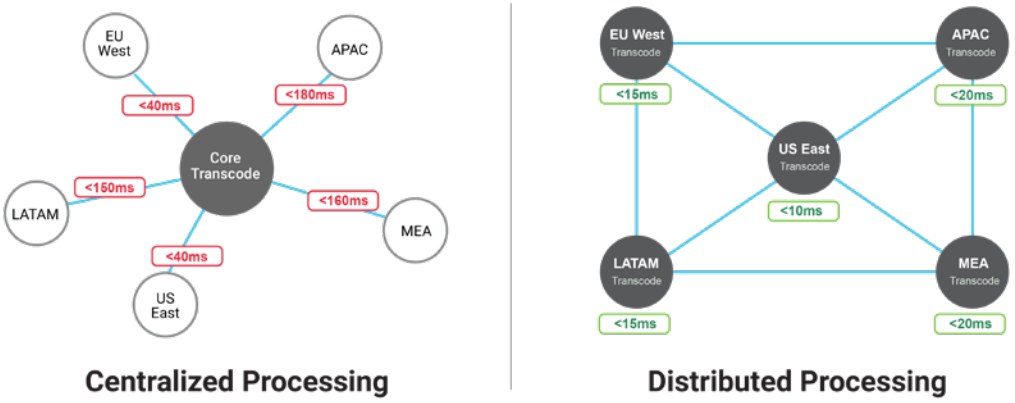
The Problem Looks Different When Latency Matters

Most conversations about video infrastructure begin with throughput and cost. How many streams can we process per dollar? How do we consolidate workloads into the fewest possible locations to simplify operations and reduce spend? For a large class of video workflows, particularly batch VOD processing and live streams where latency tolerance is measured in seconds, this centralized model works well. Scale amortizes cost, operations stay contained, and the infrastructure is easier to reason about.

But a growing category of workloads does not fit this model. Interactive streaming, cloud gaming, real-time collaboration, and regionally regulated content all share a common trait: physical distance between the user and the processing layer directly determines whether the experience works. For these workflows, the core question is not whether video can be encoded efficiently. It is whether compute can be placed close enough to users without multiplying cost and operational complexity beyond control.

This is the question i3D.net has spent years answering. As a global infrastructure provider with deep roots in gaming and interactive media, i3D.net operates a network with more than 26 Tbps network capacity across over 60 locations worldwide and 100 IXPs with 9,000 unique direct peering connections. Their infrastructure is built around a premise that most enterprise video platforms have treated as optional: that proximity matters, and that the network between the user and the processing layer is not a secondary concern but a primary design variable.

Exhibit 1: Centralized architectures trade latency for simplicity. Distributed models reverse that tradeoff.



Geography as the Dominant Design Variable

When video infrastructure must be geographically distributed, the economics and operations of scale behave differently. Instead of running a small number of very large sites, teams operate many smaller ones. Each location comes with its own constraints around power availability, physical space, cooling capacity, and local connectivity. Over provisioning becomes more expensive because idle capacity is replicated across regions rather than pooled in one place. Operational complexity grows as configuration, monitoring, and failure handling must be coordinated across a wider footprint. Inefficiencies that are manageable at a single centralized site compound quickly when repeated across ten, twenty, or fifty locations.

In these environments, architecture is shaped less by raw performance benchmarks and more by a practical question: where can infrastructure realistically exist? The Middle East, where capacity pricing can be a significant hurdle. APAC, where routing is complex and submarine cable breaks are a recurring reality. Latin America, where import dependencies add friction and cost. China, where regulatory requirements and the Great Firewall create their own layer of constraints. These are not edge cases. For any platform serving a global audience, they are the operational norm.

Transcoding as the Architectural Pivot Point

CDNs and edge caching have largely solved the locality problem for video delivery. Encoded content can be placed close to viewers with well-established tooling. But the transcoding layer, the step where raw or contribution-quality video is transformed into deliverable formats, is often still centralized by default.

For latency-sensitive workflows, this creates a structural tension. Centralized transcoding simplifies operations but introduces unavoidable delay and cross-region traffic.

KEY INSIGHT

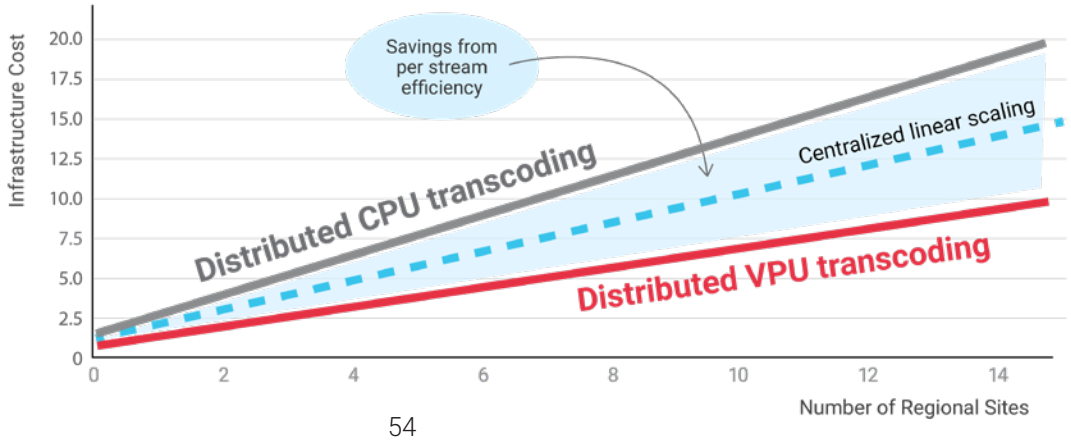
Content delivery solved the locality problem years ago through CDNs and edge caching. Transcoding has not. It remains, by default, a centralized function, and for latency-sensitive workloads, that default creates tension that delivery optimization alone cannot resolve.

Distributed transcoding reduces latency but multiplies infrastructure requirements, operational burden, and the blast radius of configuration drift.

This makes transcoding the architectural pivot point. If the transcode layer is heavy, power-hungry, or demand unpredictable, it resists distribution. The cost penalty of replicating inefficient hardware across many sites becomes prohibitive, and operational teams cannot maintain consistency at scale. If the transcode layer is dense, predictable, and efficient, it becomes movable. It can travel to the locations where latency constraints demand it without breaking the economics or the operations team.

Efficiency determines whether transcoding can be distributed without becoming prohibitive.

Exhibit 2:
In distributed systems, inefficiency does not accumulate linearly. It multiplies across every site.



What VPUs Change in Distributed Environments

Within the VPU Ecosystem, VPUs are not positioned as faster encoders. Their impact is architectural. They change where transcoding can feasibly live by addressing the three constraints that make distributed processing expensive.

First, density becomes a planning unit rather than a nice to have. Higher streams per device reduce the number of servers required at each site, which matters most when regional locations are constrained by rack space and available power. A single VPU-equipped server replacing four or five CPU-based machines at a regional site is not a performance optimization. It is the difference between a deployment that fits within the site's constraints and one that does not. Without VPUs, a service's ability to scale will be limited by operational cost and power availability.

Second, power efficiency becomes a first-order enabler. Lower watts per stream reduce the penalty of running compute in many locations simultaneously. For a centralized deployment, a 30% power reduction is a welcome cost saving. For a distributed deployment across thirty sites, that same 30% may determine whether the entire architecture is financially viable.

Third, predictability improves. When transcoding performance is deterministic, meaning the same input produces the same resource consumption regardless of content complexity, capacity planning becomes a regional exercise rather than a continuous reactive process. Teams can model per-site requirements with confidence and avoid the over provisioning that unpredictable CPU-based transcoding demands.

THE BIGGER PICTURE

In centralized environments, efficiency is primarily about saving money. In distributed, latency-sensitive environments, efficiency is about enabling architectures that would otherwise be impractical. The savings are real, but the strategic value is in the performance gains they unlock.

Where i3D.net Fits in This Picture

In the VPU Ecosystem, i3D.net addresses the placement question directly: where can video processing run when latency budgets are tight and users are geographically distributed?

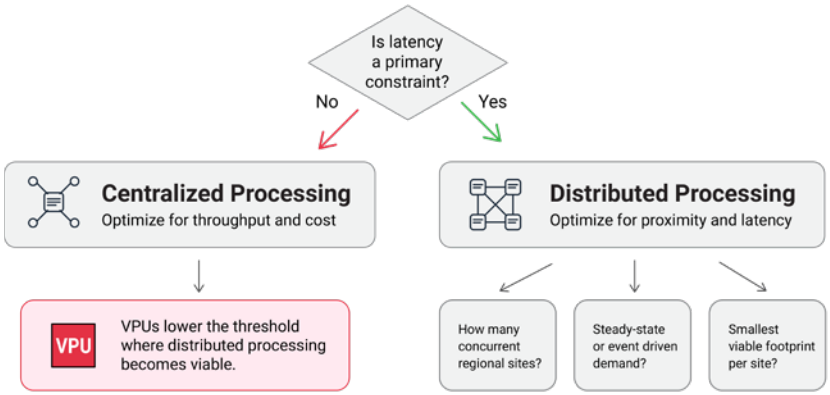
Rather than forcing all processing into a small number of centralized regions, i3D.net's infrastructure supports architectures that prioritize proximity. Video workflows can be placed closer to users, aligned with regional demand patterns, and designed around low-latency connectivity rather than distance from a core data center. For interactive streaming and gaming workloads, where latency targets of roughly 30 milliseconds or less are essential for a smooth real-time interactive or gaming experience, this proximity is not a luxury. It is a baseline requirement.

The practical value of this model extends beyond latency. A distributed architecture also changes the failure profile. When workloads run across multiple regional sites, an outage at any single location affects only a small subset of users instead of the entire global footprint. This is a meaningful resilience gain for platforms where uptime directly translates to user trust and revenue.

VPUs provide the efficiency layer that makes this model economically and operationally viable. i3D.net provides the infrastructure, the physical locations, connectivity, and operational readiness, where that efficiency translates into measurable latency and responsiveness gains.

Architectural Patterns That Become Viable

When transcoding becomes lighter and more predictable, several deployment patterns that were previously impractical become realistic options. Regional live processing, where streams are transcoded close to viewers before delivery, avoids the round-trip penalty of sending contribution feeds to a distant core site and back. Event-driven capacity, where regional sites scale briefly to meet demand spikes without maintaining large idle fleets, becomes feasible when the per-site footprint is small enough to spin up quickly. Hybrid models, where steady-state processing runs locally and overflow spills to centralized regions during peak demand, offer a middle path between full distribution and full centralization.



These patterns are not theoretical. They reflect the operational realities of platforms that serve global audiences with latency-sensitive content. The common thread is that each pattern depends on the transcode layer being efficient enough to deploy at scale across many locations without overwhelming either the budget or the operations team.

Exhibit 3: The placement decision starts with latency constraints and flows into site-level feasibility questions.

Economics at the Edge

In centralized systems, inefficiency accumulates linearly. One site runs at 60% utilization, and the waste is contained. In distributed systems, that same inefficiency multiplies. Costs are evaluated per site rather than per platform. Idle capacity is replicated across regions. Power waste compounds geographically rather than averaging out.

This makes per-stream efficiency disproportionately valuable in distributed architectures. A modest efficiency improvement at a single centralized site is a line item. The same improvement applied across dozens of regional locations becomes a significant aggregate gain, often enough to shift a deployment from marginally viable to clearly sustainable.

VPUs reduce what might be called the replication penalty of distributed infrastructure. By lowering the per-site cost in power, space, and hardware, they lower the threshold at which regional processing makes financial sense. For platforms considering expansion into regions that were previously too expensive to serve locally, this is the efficiency that enables reach.

Operational Reality: What Still Matters

Efficiency does not remove operational complexity. It changes the threshold at which complexity becomes manageable.

Distributed video systems still require careful orchestration, regional failover strategies, and observability at per-site granularity. Automation becomes essential to prevent configuration drift and operational overhead from overwhelming teams. These challenges do not disappear because the transcode layer is more efficient.

Two operational realities deserve particular attention. First, proximity to users does not eliminate the need for rigorous network validation. The internet remains inherently dynamic, and physical distance is not a reliable proxy for performance. Regional carriers may lack optimal routes, peering may be inconsistent, and congestion can shift unpredictably. Teams must continuously verify real-world metrics, including latency, jitter, packet loss, and throughput, to ensure that local deployments truly deliver local-grade performance.

Second, regional regulations introduce their own layer of friction. Import and export controls, certification requirements, and evolving frameworks like NIS2 can slow or restrict the movement of equipment, constraining how

quickly infrastructure can be expanded or rebalanced. These factors become part of the operational landscape, requiring early awareness and deliberate planning.

What efficiency provides is margin. With fewer nodes, lower power draw, and more predictable performance per site, operational challenges become tractable rather than dominant. The team's energy shifts from managing infrastructure constraints to optimizing the service running on top of it.

Closing Perspective

The question at the center of this article is not new, but it is becoming more urgent: where should video compute live when geography, user proximity, and latency constraints dominate system design?

The traditional answer has been to centralize. Build large, efficient processing hubs and accept the latency tradeoff. For many workloads, that remains the right answer for traditional video platforms. But for a growing category of real-time, and regionally sensitive applications, like cloud gaming and interactive video applications, the tradeoff is no longer acceptable.

VPUs make transcoding light enough to move anywhere the service and quality targets require. i3D.net provides the global infrastructure where that mobility creates real user value. Together, they shift the conversation from a defensive question, "Can we afford to run video close to users?", to a strategic one: "Where does it make the most sense to run our video operations?"

THIS ARTICLE IS PART OF THE VPU ECOSYSTEM SERIES EXPLORING HOW PURPOSE-BUILT SILICON IS RESHAPING VIDEO INFRASTRUCTURE ACROSS THE TECHNOLOGY STACK. EACH ARTICLE EXAMINES A DIFFERENT ARCHITECTURAL PERSPECTIVE THROUGH THE LENS OF ONE PARTNER WITHIN THE ECOSYSTEM.

8K

<300 ms

Glass-to-glass. 8K60p. Over standard internet.

Live 8K used to require a truck and a satellite. Not anymore.

QIKENC8K MAX delivers broadcast-quality 8K over standard internet. WebRTC. No VPN. Powered by NETINT VPU technology.

MISAO Network

misao.co.jp

CASE STUDY: TMPS | MISAO NETWORK | NETINT

When Milliseconds Define the Experience

How TMPS and Misao Network Used NETINT VPUs to Deliver Ultra-Low-Latency 8K Video over WebRTC.

There is a class of video experience where traditional streaming simply does not work. Not because it delivers poor quality, but because it delivers that quality too late. Live sports production, interactive broadcast, remote surgical monitoring, real-time surveillance: these are domains where the gap between what happened and what appears on screen must be measured in tens of milliseconds, not seconds.

TMPS, a Japanese systems integrator specializing in professional video workflows, set out to explore whether this kind of performance was achievable at 8K resolution. The goal was ambitious: deliver broadcast-quality 8K video with sub-100 millisecond glass-to-glass latency, using a workflow stable enough for commercial deployment. The team partnered with Misao Network, whose expertise in WebRTC transport optimization addressed the delivery side of the equation, and selected NETINT VPUs to handle real-time encoding without the latency variability that CPU-based compression introduces.

The resulting system was showcased at InterBEE in Japan and subsequently adapted for demonstration at the 2026 NAB show in the VPU Ecosystem booth. This case study examines the architecture, design decisions, and practical tradeoffs behind that workflow.

The Challenge: Three Constraints, One System

High-resolution live video presents a fundamental tension. As resolution and frame rate increase, so do encoding complexity, bandwidth requirements, and latency sensitivity. These three factors pull in different directions, and traditional streaming architectures resolve the tension by sacrificing one of them. HLS and DASH prioritize quality and reliability but accept latencies that can range up to 30 seconds or more. RTMP reduces delay but struggles with modern codecs and high resolutions. WebRTC achieves sub-second latency but has historically been limited to lower resolutions and simpler encoding profiles.

TMPS wanted to know whether it was possible to satisfy all three requirements simultaneously: 8K video at broadcast-quality fidelity, ultra-low latency suitable for interactive or real-time viewing, and operational predictability required for live production environments. The answer, they believed, depended on treating encoding, transport, and system design as a single integrated problem rather than three separate ones.

KEY INSIGHT

The challenge was not simply compressing 8K video quickly. It was maintaining deterministic, predictable behavior across every stage of the pipeline, from capture to display, so that latency could be budgeted and controlled rather than merely hoped for.

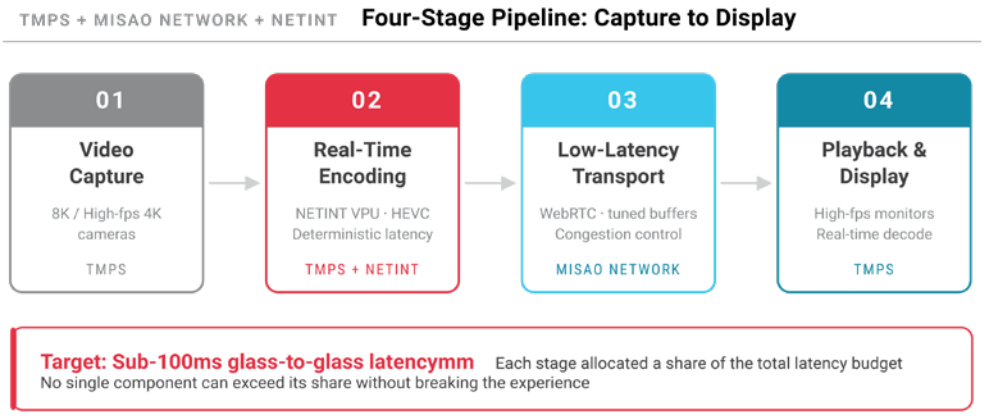


Exhibit 1: The four-stage pipeline from capture to display, with each partner contributing specialized capability.

System Architecture: Four Stages, Three Partners

The workflow that TMPS designed consists of four primary stages, each addressing a specific portion of the end-to-end latency budget. The architecture reflects a deliberate decision to assign each challenge to specialized technology rather than asking any single component to solve everything.

In the first stage, high-resolution video is captured using professional or DSLR-class cameras capable of 8K or high-frame-rate 4K output, depending on the desired configuration. The captured signal is passed directly to NETINT VPUs for real-time HEVC encoding. The encoded bitstream is then handed to Misao Network's WebRTC transport layer, which handles packetization, buffer management, congestion control, and delivery. Finally, the stream is decoded and displayed on high-frame-rate monitors where viewers can observe motion smoothness, responsiveness, and image detail.

Each stage was designed with a specific latency allocation. The total budget of sub-100 milliseconds glass-to-glass meant that no single stage could consume more than its share without breaking the entire experience.

TMPS + MISAO NETWORK + NETINT **Workflow Stages And Partner Responsibilities**

STAGE	IMPLEMENTATION	RESPONSIBILITY
01 - Video Capture	Professional / DSLR cameras. 8K or high-fps 4K output	TMPS
02 - Real-Time Encoding	NETINT VPUs · HEVC codec. Deterministic, bounded latency	TMPS + NETINT
03 - Low-Latency Transport	WebRTC · tuned buffers. Congestion control · packetization	MISAO NETWORK
04 - Playback & Display	High-frame-rate monitors. Real-time decode and verification	TMPS

Why VPUs Were Critical to the Design

Encoding high-resolution, high-frame-rate video in real time introduces several risks in live environments. CPU-based software encoding, while flexible, produces variable encoding latency depending on content complexity, system load, and thermal conditions. A frame of simple content might encode in 8 milliseconds while a complex scene takes 30. That variability cascades downstream: the transport layer cannot optimize buffer sizes when the input timing is unpredictable, and the display stage cannot maintain smooth playback when frames arrive at irregular intervals.

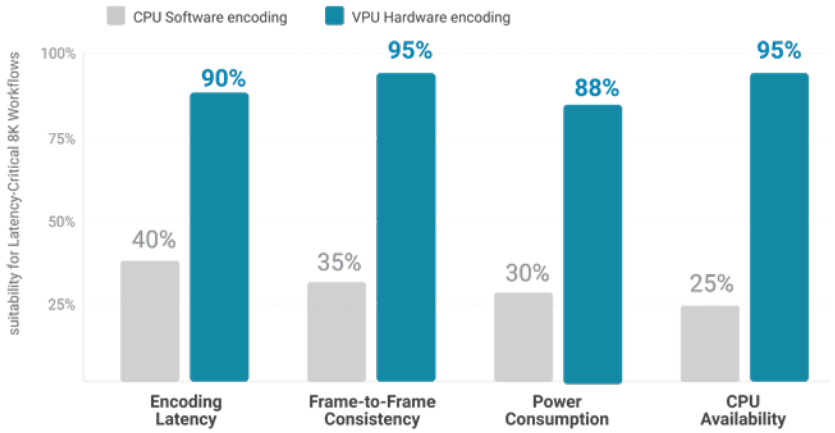


Exhibit 2: VPU hardware encoding outperforms CPU software encoding across every dimension critical to latency-sensitive workflows.

NETINT VPUs addressed this challenge through three characteristics that proved essential to the workflow. First, deterministic encoding behavior: the hardware processes each frame in a consistent, bounded time regardless of content complexity, which is critical for latency budgeting. Second, high density per device, which reduced system complexity by eliminating the need for multi-server encoding clusters. Third, bounded power consumption, which was important for the live demonstration environments where the system needed to operate for extended periods without thermal throttling.

By isolating video processing onto dedicated hardware, TMPS was able to focus on transport and workflow optimization without compensating for encoder variability. The VPU became the fixed, predictable element in the pipeline that allowed everything else to be tuned around it.

WebRTC Optimization: Misao Network's Transport Layer

Misao Network's contribution focused on the transport problem: how to move a high-bitrate encoded stream from encoder to decoder with minimal added delay. WebRTC was selected as the transport protocol because of its native support for real-time delivery, built-in congestion control, and sub-second latency characteristics. But pushing WebRTC to handle 8K video at production quality required significant optimization beyond its standard configuration.

THE BIGGER PICTURE

In latency-critical systems, predictability is more valuable than peak performance. A slower encoder that delivers every frame in exactly 10 milliseconds is more useful than a faster one that averages 8 milliseconds but occasionally spikes to 25. The TMPS workflow depended on this consistency to maintain its sub-100 ms target across sustained operation.

The key areas of tuning included codec and packetization choices suitable for the high bitrates that 8K HEVC generates, buffer sizes reduced to minimize end-to-end delay without introducing jitter artifacts, jitter and network variability management that maintained visual quality under real-world conditions, and bitrate strategies balanced against WebRTC's congestion control algorithms, which are designed for conversational video and needed adaptation for high-resolution production streams.

The combination of predictable encoding latency from the VPUs and carefully tuned WebRTC transport produced what the team described as a 'mirror-like' viewing experience, where motion on screen appeared immediate rather than delayed. This responsiveness was directly observable during the public demonstrations: viewers could see camera movement reflected on the display monitors with no perceptible lag.

The Latency Budget: Where Every Millisecond Goes

Achieving sub-100 millisecond glass-to-glass latency required a disciplined approach to latency allocation across every stage of the pipeline. In traditional streaming architectures, latency accumulates through encoding buffers, segmentation delays, CDN propagation, and player buffering, with each stage adding hundreds of milliseconds or more. The TMPS workflow eliminated most of these sources by replacing segment-based delivery with frame-level WebRTC transport and replacing variable CPU encoding with deterministic VPU processing.

The practical implication is significant. Traditional live streaming protocols like HLS typically operate with 6 to 30 seconds of end-to-end latency. Low-latency HLS reduces this to 2 to 5 seconds. Standard WebRTC for conversational video achieves 200 to 500 milliseconds. The TMPS workflow targeted a range below even standard WebRTC, pushing into territory typically associated with dedicated hardware video links rather than software-based streaming systems.

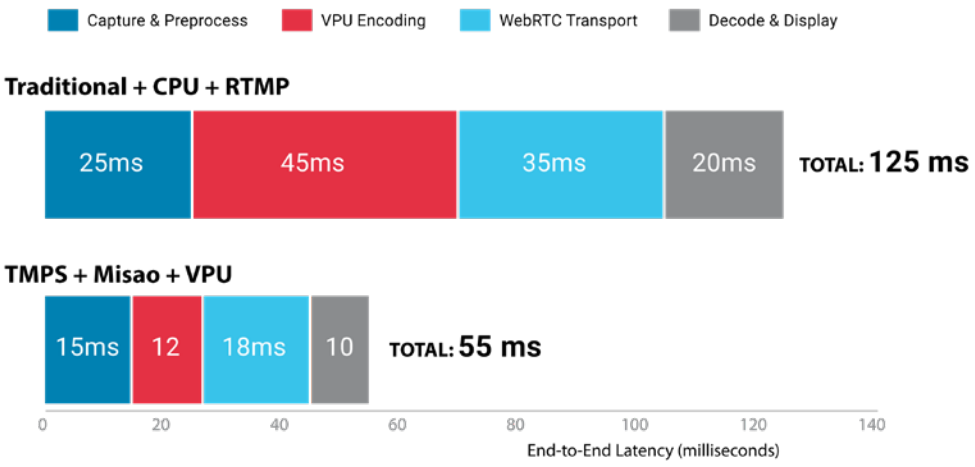
Demonstration Results and Practical Lessons

During public demonstrations, the system achieved visually smooth playback at high resolution, extremely low perceived latency, and stable operation over extended demonstration periods. The specific resolution and frame rate varied based on display availability and venue constraints, but the underlying workflow proved adaptable across configurations without requiring fundamental redesign. Several practical lessons emerged from the project that carry implications beyond this specific workflow.

1. Display technology matters more than expected. Monitor refresh rate and HDMI capabilities meaningfully affected perceived latency. The team discovered that display-side bottlenecks could undermine encoding and transport optimizations if not addressed, reinforcing the importance of treating the entire pipeline as a single system.

2. Predictability beats peak performance. Consistent behavior under sustained load proved

Exhibit 3: Latency breakdown comparing a traditional pipeline with the optimized workflow of TMPS + Misao + VPU.



more valuable than headline throughput numbers. A system that maintained stable latency for eight hours of continuous demonstration was more useful than one that achieved lower latency in short bursts but degraded over time.

3. Live environments amplify complexity. Hardware stability and simplicity reduced operational risk during public showcases. Every additional component in the signal chain represented both a potential failure point and a potential latency contribution. The team's decision to use VPUs rather than multi-server CPU encoding clusters eliminated an entire category of operational complexity.

What This Workflow Demonstrates

The TMPS + Misao Network collaboration demonstrates that ultra-low-latency, high-resolution video delivery is achievable when encoding, transport, and operational constraints are treated as a single integrated system. By combining Misao Network's WebRTC expertise with NETINT's dedicated video-processing hardware, the team was able to push resolution and frame rate boundaries while maintaining the responsiveness required for live demonstrations and production-adjacent use cases.

KEY INSIGHT

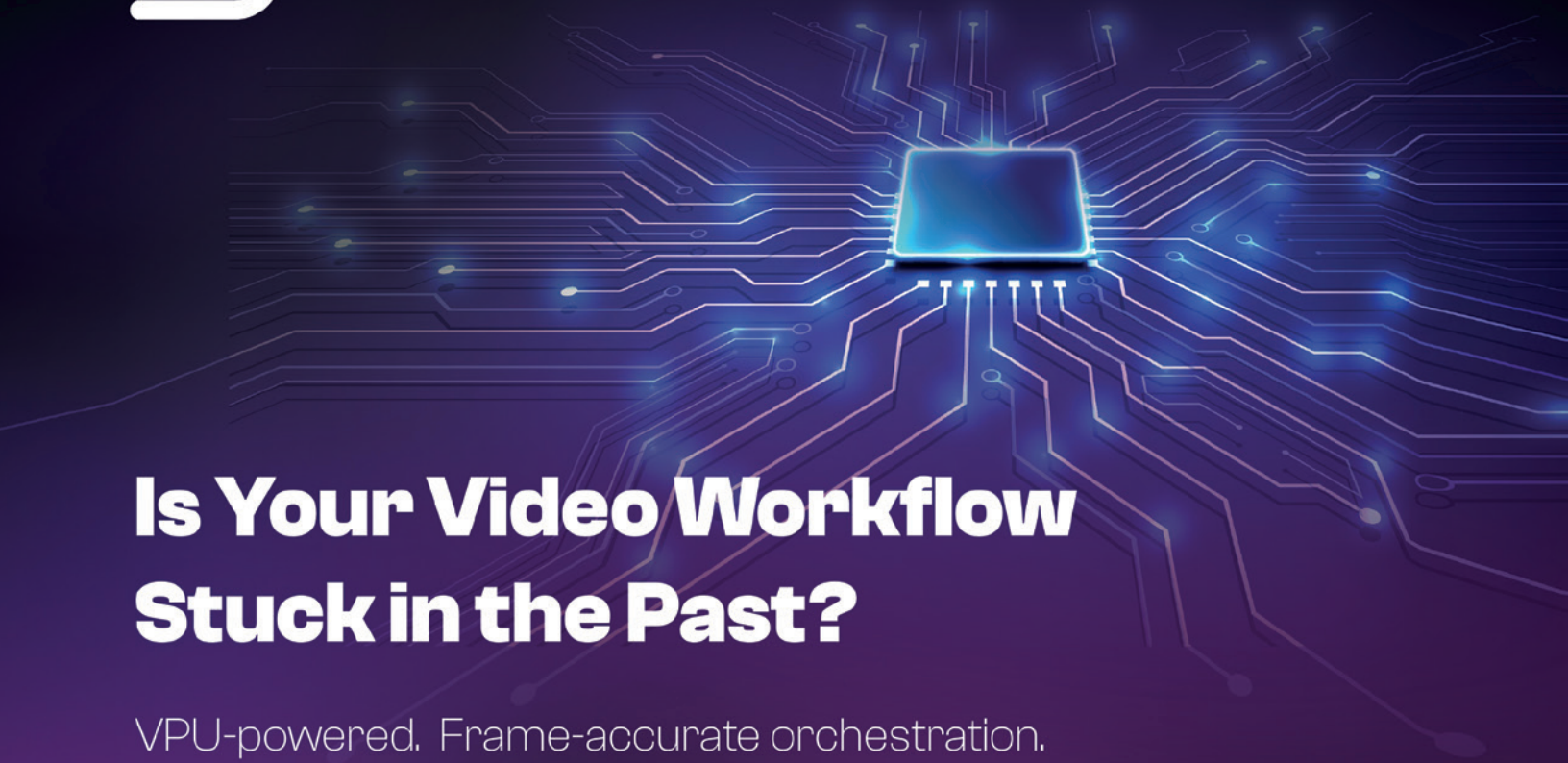
The workflow was designed to be adaptable. Rather than optimizing for a single fixed configuration, the architecture allowed the team to adjust resolution, frame rate, and encoding parameters to match venue conditions without rebuilding the system.

The project also illustrates a broader shift in how professional video workflows are being constructed. Rather than relying on monolithic, proprietary solutions, the team assembled a pipeline from specialized components: purpose-built encoding silicon, optimized transport software, and careful system integration. Each component does what it does best, and the system's performance emerges from how those components interact rather than from any single technology alone.

For organizations exploring real-time video delivery at high resolution, the TMPS workflow offers a practical reference architecture. The encoding predictability of VPUs, the latency characteristics of tuned WebRTC transport, and the systems integration expertise to bring them together represent a template that extends well beyond trade show demonstrations into live production, remote collaboration, and interactive broadcast.

THIS ARTICLE IS PART OF THE VPU ECOSYSTEM SERIES EXPLORING HOW PURPOSE-BUILT SILICON IS RESHAPING VIDEO INFRASTRUCTURE ACROSS THE TECHNOLOGY STACK. EACH ARTICLE EXAMINES A DIFFERENT ARCHITECTURAL PERSPECTIVE THROUGH THE LENS OF ONE PARTNER WITHIN THE ECOSYSTEM.

SCALSTRM



Is Your Video Workflow Stuck in the Past?

VPU-powered. Frame-accurate orchestration. On-demand by design. Our software is built for the economics of streaming at scale.

50%
Savings per channel

75%
Reduction in server footprint

80%
Lower energy consumption

90%
Less storage and compute with Just-In-Time Transcoding

NAB 2026 → **Booth W2713**

www.scalstrm.com

SCALSTRM

VIDEO ORCHESTRATION

The Control Plane Imperative: Operating Video at Scale

Why Orchestration, Not Raw Speed, Determines Whether Video Infrastructure Scales With Confidence.

The Operational Breaking Point

Large video systems rarely fail because encoding stops working. They fail because operational complexity overtakes human and system capacity. The technical components keep running, but the ability to reason about, direct, and recover the whole system erodes as concurrency grows.

As pipelines scale, teams contend with more simultaneous jobs, more pipeline stages, more regions, and more failure domains. State proliferates. Dependencies multiply. Small misalignments propagate quickly, surfacing as symptoms far from their root causes.

This is the paradox of modern video infrastructure: the hardware has never been more capable, yet the systems it powers have never been harder to operate. The bottleneck has shifted from processing power to operational control. At scale, control is the system that matters.

When Efficiency Amplifies Complexity

Video Processing Units (VPUs) have materially improved the economics of transcoding. They increase density, reduce power consumption per stream, and make capacity more predictable. These gains are significant, but they also reshape the operational landscape in ways that are easy to underestimate.

When transcoding becomes cheaper and denser, teams respond by running more concurrent jobs, distributing workloads more aggressively, and concentrating more work on fewer resources. Each of these decisions is rational in isolation. Together, they expand the blast radius of any single mistake.

Efficiency, in other words, does not remove the operational problem. It shifts it. The faster and denser the system becomes, the more critical it is to decide precisely where jobs run, when they run, how they fail, and how they recover. Without those decisions being made deliberately, efficiency becomes a liability rather than an advantage.

KEY INSIGHT

VPU efficiency is a necessary condition for scalable video infrastructure, but it is not sufficient. Efficiency without orchestration concentrates risk. Orchestration is what converts hardware gains into system reliability.

The Hidden Pipeline

Modern video pipelines are no longer linear chains of discrete tasks. They are event-driven, multi-stage systems with dynamic, often unpredictable, behavior.

Ingest triggers processing. Processing fans out across encoding ladders and codecs. Outputs feed packaging and delivery. Monitoring loops feed back into scheduling, scaling, and retry logic. Each stage operates semi-independently, and each introduces its own failure modes, timing constraints, and resource requirements.

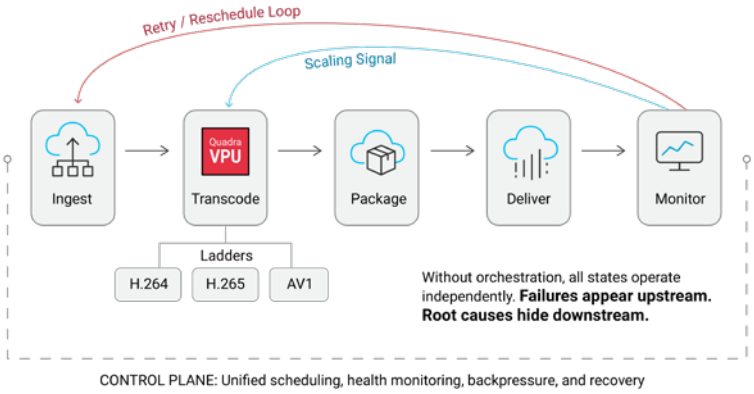


Exhibit 1: A modern video pipeline viewed as an event-driven system with control plane coordination.

Newer architectures add further dynamism. Just-In-Time transcoding, for example, defers encoding until a viewer actually requests a stream, creating segments on demand rather than pre-encoding an entire catalog. Scalstrm's JIT implementation supports H.264, HEVC, and AV1 in software and hardware (VPU), and VP9 in software, generating adaptive bitrate ladders in real time. The approach can reduce storage and compute consumption by up to 90 percent, but it also means that the pipeline's workload is inherently unpredictable, shaped by viewer behavior rather than batch schedules.

The result is a pipeline that looks manageable on a whiteboard but behaves unpredictably under load. Failures appear upstream while their root causes hide downstream. Operators add compute to relieve congestion, only to discover they have made it worse by overwhelming a bottleneck further along the chain.

Orchestration as the Leverage Point

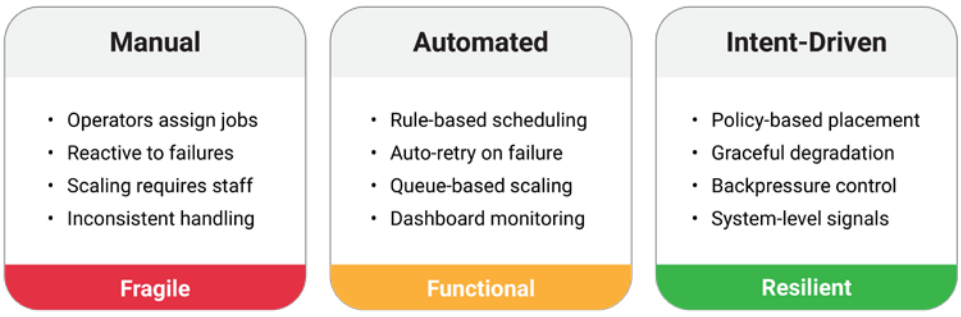
Once video infrastructure incorporates specialized compute such as VPUs, orchestration determines whether that efficiency is realized or wasted.

A capable control plane allows teams to allocate VPU capacity intentionally rather than opportunistically, balance workloads across heterogeneous resources, prevent overcommitment, and isolate failures before they cascade. It transforms scheduling from a mechanical task into a strategic function.

In practice, several operational patterns emerge when orchestration is treated as a first-class system component rather than an afterthought. Intent-based scheduling places workloads according to policy, not just availability. Failure isolation ensures that individual pipeline stages can fail without collapsing the entire workflow. Capacity transparency gives teams real-time understanding of not just how much capacity exists, but precisely how it is being consumed.

These patterns are difficult to achieve through scripting or manual intervention. They require continuous coordination across pipeline stages, informed by real-time system state rather than static assumptions.

Exhibit 2: The orchestration maturity model, from fragile manual control to resilient intent-driven automation.



Where Scalstrm Fits

Scalstrm is a Stockholm-based streaming infrastructure company founded in 2017 by Ola Bengtsson and Patrik Alm, who bring 75 years of combined experience in media technology. The company now operates across 35 countries with 57 partner organizations, and its platform is used by operators including Telia (TV4, C More), Altibox, Tele2, and NEP Group.

Within the VPU Ecosystem, Scalstrm addresses the coordination problem directly. Rather than treating encoding, packaging, and delivery as isolated steps, the platform unifies them into four tightly integrated components: an Origin Platform for cloud-based tile recording and live stream management, a CDN Platform with a real-time programmable HTTP pipeline for content delivery, the Just-In-Time Transcoding engine described earlier, and an Origin Shield that provides content-aware caching across RAM, NVMe, and disk tiers.

The architecture is built on what Scalstrm calls a scalable micro-component model. Each component can be deployed and scaled independently across on-premises, private cloud, public cloud, or hybrid environments. This modularity matters because it allows the control plane to manage heterogeneous infrastructure as a single system rather than a collection of disconnected services.

At the hardware level, the numbers are concrete. A single server running Scalstrm's Origin Platform can manage up to 800 adaptive bitrate channels while pushing 200 Gbps of egress. The transcoding layer integrates with CPU, GPU, and dense VPU architectures, including NETINT Quadra T1U units, supporting HEVC encoding across SD, HD, and 4K resolutions. New live channels can be launched in seconds through instant scaling, with operators paying only for active capacity.

Decisions about placement, retries, and prioritization are informed by real-time system state. When a VPU node reaches its encoder session limit, the control plane deprioritizes it and redistributes work. When a downstream packaging service slows, back pressure is applied automatically to upstream stages rather than allowing queues to grow unchecked. The Origin Shield layer adds a further coordination point, automatically identifying and protecting live streams, catch-up TV, and network PVR content with redundancy for uninterrupted service.

Scalstrm treats all compute resources as abstract execution targets, exposing hardware capabilities through normalized metrics. This allows scheduling decisions to be made based on actual performance characteristics rather than static labels, keeping orchestration hardware-agnostic while still optimizing for throughput, energy efficiency, and cost per processed hour.

The Failure Equation

Even with strong orchestration, video systems remain complex. Teams must still manage imperfect observability, evolving workload characteristics, and external dependencies such as networks and third-party services. These challenges do not disappear.

What changes is how failures are experienced. Two failure modes are common and consistently underestimated. The first is silent performance degradation from partial resource exhaustion. Unlike complete node failures, conditions such as VPU encoder session limits being reached, CPU thermal throttling under sustained load, or memory pressure causing increased I/O wait don't trigger

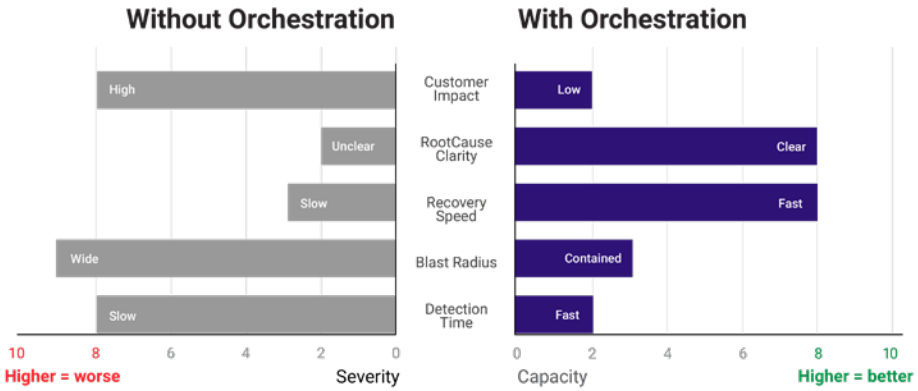


Exhibit 3:
How orchestration transforms failure response across five critical dimensions.

immediate alarms. In un-orchestrated pipelines, jobs continue to flow to affected nodes, processing times drift upward, and SLAs are missed without a clear root cause. Operators often misdiagnose the issue as bad content or temporary load.

The second is cascading backlogs caused by downstream bottlenecks. When packaging services, storage endpoints, or CDN ingest connections slow down, an un-orchestrated system keeps producing output at full speed. Queues grow invisibly. Failures surface upstream, far from the real problem. Operators respond by adding more compute, which only deepens the congestion.

With coordinated control, both patterns are handled differently. Scalstrm's platform continuously monitors resource health through its WebUI dashboard, providing real-time metrics and analytics across every pipeline stage. Degraded nodes are automatically deprioritized, and back pressure is applied when downstream systems falter. The auto-scalable architecture dynamically allocates resources based on real-time traffic, eliminating the need for operators to guess at capacity during demand spikes. Instead of cascading outages, teams encounter bounded incidents with clearer signals, controlled impact, and defined recovery paths.

THE BIGGER PICTURE

Orchestration does not eliminate failures. It converts unknown unknowns into known, bounded risks. The distinction between feeling brittle and feeling resilient is not about the absence of problems. It is about the presence of predictable, controlled responses.

Economics Follow Control

Operational efficiency and economic efficiency are inseparable in video infrastructure.

Without orchestration, capacity is stranded across underutilized nodes, peaks are over provisioned because there is no mechanism to redistribute load dynamically, and failures generate hidden costs through retries and reprocessing. Efficiency gains from VPU hardware erode quietly, diluted by the operational overhead of running an uncoordinated system.

With coordinated control, the gains compound. Scalstrm reports that its live transcoding layer delivers 50 percent savings per channel compared to conventional architectures, while requiring 75 percent less server footprint and 80 percent less power consumption. These numbers reflect what becomes possible when orchestration ensures that workloads are placed where capacity actually exists, rather than distributed across partially utilized or thermally constrained nodes.

The sustainability dimension is increasingly difficult to ignore. At 80 percent lower energy consumption and a corresponding reduction in carbon footprint, the economic case and the environmental case converge. Organizations operating at scale cannot separate cost-per-stream from watts-per-stream indefinitely.

In this sense, orchestration is not an operational luxury. It is an economic multiplier. The return on investment in VPU hardware is directly proportional to the quality of the control plane that manages it.

Four Questions That Reveal Readiness

Teams evaluating their readiness to operate VPU-accelerated video pipelines at scale should begin with four questions. If these questions prove difficult to answer, efficiency gains alone will not translate into operational confidence.

First: can you see where every job is running and why? If the answer requires checking multiple dashboards, querying separate systems, or asking a colleague, the pipeline lacks the visibility that coordinated control provides.

Second: can you change scheduling behavior without redeploying the pipeline? If policy changes require code modifications and release cycles, the system cannot respond to operational realities at the speed they demand.

Third: can individual stages fail without failing the workflow? If a single component failure triggers cascading impact across the entire pipeline, the system lacks the isolation that resilient architectures require.

Fourth: can you explain capacity usage in real time? If utilization is only understood in retrospect, through post-incident analysis or monthly reports, the team is operating reactively rather than proactively.

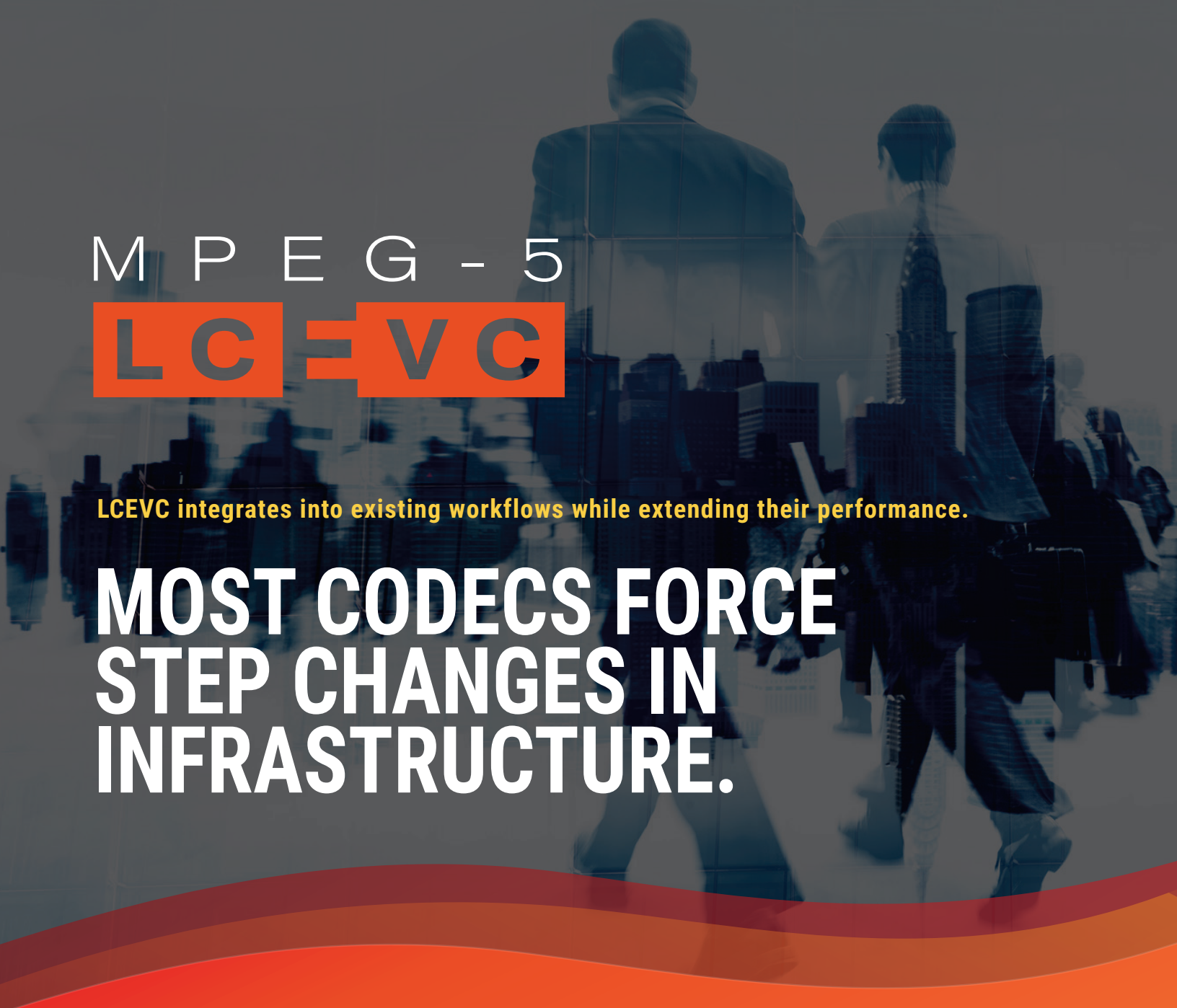
Closing Perspective

As video infrastructure evolves, performance improvements increasingly come from how systems are controlled, not just how fast individual components run.

VPUs make video processing efficient. Platforms like Scalstrm make that efficiency usable. Together, they represent a shift in how the industry thinks about scale: not as a problem of raw throughput, but as a problem of coordination, visibility, and control. The fact that a platform born in Stockholm now orchestrates video infrastructure across 35 countries suggests the market has arrived at the same conclusion.

The organizations that navigate this transition successfully will be those that recognize orchestration not as an operational add-on, but as the foundational layer that determines whether their infrastructure scales with confidence or collapses under its own complexity.

THIS ARTICLE IS PART OF THE VPU ECOSYSTEM SERIES EXPLORING HOW PURPOSE-BUILT SILICON IS RESHAPING VIDEO INFRASTRUCTURE ACROSS THE TECHNOLOGY STACK. EACH ARTICLE EXAMINES A DIFFERENT ARCHITECTURAL PERSPECTIVE THROUGH THE LENS OF ONE PARTNER WITHIN THE ECOSYSTEM.



MPEG-5 LCEVC

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

MPEG-5 LCEVC

CODEC TECHNOLOGY

When Standards Become Strategy

How MPEG-5 LCEVC Is Rewriting the Economics of Video Quality.

The video industry has spent decades locked in a familiar cycle: a new codec standard is ratified, hardware support lags by years, encoding infrastructure must be rebuilt, and the ecosystem gradually migrates typically over five to seven years. But the gap between laboratory benchmarks and operational reality persists, and most platforms still operate well below the theoretical efficiency frontier.

MPEG-5 LCEVC (Low Complexity Enhancement Video Coding) breaks this pattern. Rather than replacing existing codecs, it enhances them. Standardized by ISO/IEC as Part 2 of the MPEG-5 family, LCEVC operates as an enhancement layer that pairs with any base codec from legacy H.264/AVC through HEVC and AV1 to the newest H.266/VVC.

The strategic significance lies not in compression efficiency alone, but in what that efficiency enables: integration into existing encoding pipelines, decoding on installed devices through firmware updates, and incremental adoption without wholesale infrastructure replacement.

The Evidence Base

Claims about codec performance are easy to make and difficult to verify. What distinguishes LCEVC is the breadth of independent validation behind it. The foundational assessment came from MPEG itself. In 2021, formal verification tests at the 134th MPEG meeting used ITU-R BT.500 DSIS methodology, the gold standard for subjective quality

STUDY / SOURCE	BASE CODEC	SAVINGS	METHODOLOGY
MPEG Verification Test (2021)	H.264/AVC	**46% (UHD)**	ITU-R BT.500 DSIS subjective MOS
MPEG Verification Test (2021)	HEVC	**31% (UHD)**	ITU-R BT.500 DSIS subjective MOS
Frontiers in Signal Processing	VVC	**52–55%**	4K HDR ITU-R BT.500 DSIS (30 viewers)
ACM Mile High Video (2022)	x264 / x265	**42% / 39%**	VMAF objective + subjective (gaming)
Springer Multimedia Tools (2023)	AVC / HEVC	**39% / 21%**	VMAF under packet loss conditions
NVIDIA + V-Nova (NAB 2024)	NVENC HEVC	**30%**	Low-latency cloud gaming scenarios
SPIE + V-Nova + Intel + Meta (2022)	SVT-AV1	**~40%**	Convex hull VMAF_NEG + battery tests
IEEE TCSVT Overview (2022) Univ. of Brasilia / SBTVD (2024)	AVC / HEVC / VVC VVC	**28–34% (UHD)** **36% (4K HDR)**	BD-rate: VMAF, MS-SSIM, PSNR ITU-R BT.500 DSIS subjective (30 viewers)

Sources: MPEG WG04 (2021) · Frontiers Signal Processing · SPIE (2022) · ACM Mile-High Video (2022) · IEEE TCSVT (2022) • Springer MTA (2023) · NVIDIA/V-Nova (2024) · Univ. Brasilia/SBTVD (2024)

Table 1: Summary of Independent LCEVC Research Validation.

assessment, with trained viewers in controlled laboratory conditions. The results: 46% bitrate savings when enhancing H.264/AVC and 31% when enhancing HEVC, both for UHD content. Nine independent studies have since reinforced those findings across different codecs, content types, and methodologies:

Three findings stand out. First, the consistency of results across independent labs, base codecs, and content types from broadcast television to live gaming suggests the efficiency gains are structural, not circumstantial. Second, LCEVC adds easily manageable processing overhead while enabling encoding speeds 3.1–3.6x faster than full-resolution alternatives. Third, the gains hold under real-world conditions including packet loss and low-latency constraints.

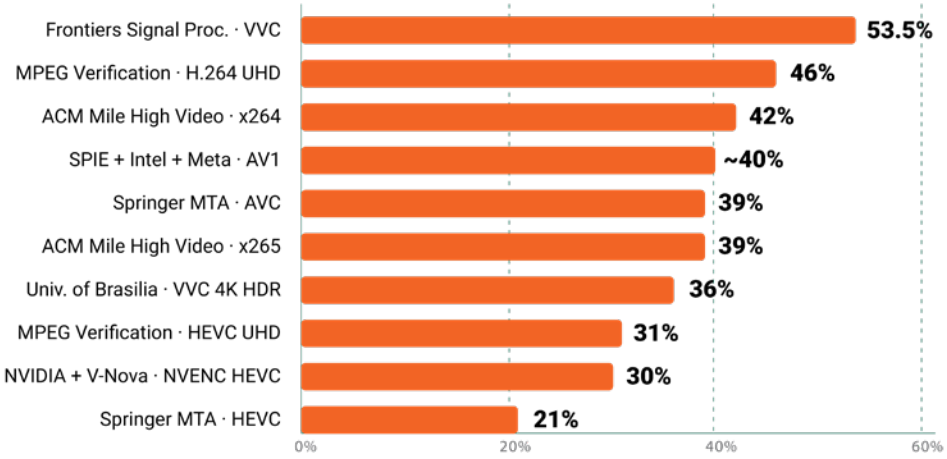


Exhibit 1: LCEVC bitrate savings by base codec, drawn from MPEG verification tests, peer-reviewed studies, and independent validation.

The Brazil Proof Point

Brazil's TV 3.0 initiative branded DTV+ for consumers provides the most consequential deployment validation. On August 27, 2025, President Lula signed the decree formalizing TV 3.0 into law. The standard mandates VVC compression, MPEG-H audio, and MPEG-5 LCEVC enhancement. It is the first national broadcast standard to require LCEVC.

The rationale was pragmatic. Brazil must deliver UHD to over 200 million viewers using limited terrestrial spectrum. Current HD services consume approximately 14 Mbps per channel on private television spectrum. By combining VVC with LCEVC, TV 3.0 achieves 4K HDR at under 10 Mbps per channel. Independent verification by the University of Brasilia confirmed that LCEVC-enhanced VVC matched subjective quality at 8.08 Mbps where standalone VVC required 12.54 Mbps.

During the Paris 2024 Olympics, Globo ran the full TV 3.0 stack in live production: UHD feeds encoded at 10 Mbps, decoded on consumer devices, including TVs from Hisense and set-top boxes with silicon from Realtek and Amlogic. Pilot DTV+ transmissions launched in Rio de Janeiro in 2025, with commercial services targeting the 2026 World Cup.

KEY INSIGHT

Brazil solved the chicken-and-egg problem that historically blocks video technology adoption by tying deployment to a concrete commercial milestone, forcing silicon vendors, encoder manufacturers, and device makers to deliver simultaneously. LCEVC's inclusion validates not just efficiency, but ecosystem maturity.

Exhibit 2: LCEVC deployment opportunities span broadcast, streaming, gaming, and pay TV sectors.

Four Deployment Vectors

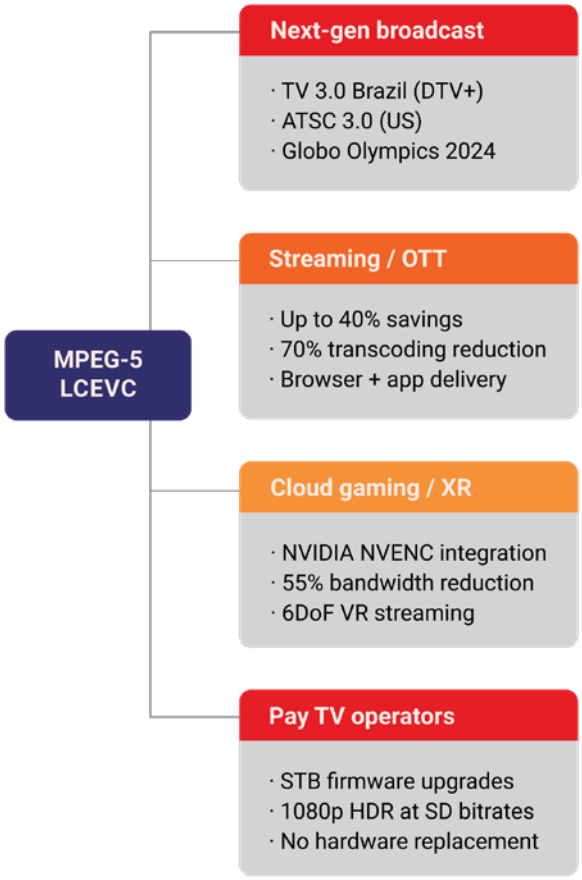
While broadcast provides the highest-profile proof point, LCEVC's codec-agnostic design makes it relevant across the full delivery spectrum.

1. For streaming and OTT platforms, LCEVC offers up to 40% compression efficiency improvement and a 70% reduction in transcoding costs and energy consumption. Research published in SPIE Proceedings demonstrated that LCEVC-enhanced SVT-AV1 achieves 75–85% encoding time savings through fast parameter selection, while maintaining equivalent quality as measured by VMAF_NEG. On the decoding side, LCEVC reduces CPU utilization by 19–51% depending on preset and quality level. These are not marginal operational savings.

2. For platforms processing millions of hours of content, the economic impact of reducing per-title encoding costs while simultaneously improving delivered quality represents a structural advantage. The technology integrates with standard encoding pipelines through commercially available MainConcept SDK, Harmonic XOS and Ateme TITAN, and decodes through widely adopted players including Shaka Player and GStreamer 1.26+.

3. For pay TV operators, the value proposition is preservation rather than replacement. Operators can deliver 1080p HDR at the bitrates currently used for SD content on their installed set-top boxes. Devices built on Realtek, Amlogic, and Montage LZ silicon support LCEVC playback through a compact software module that utilizes the existing SoC. Legacy devices continue to decode the base stream normally, while modern units gain enhanced quality through a firmware update. No truck rolls. No hardware swaps. No subscriber disruption.

4. For cloud gaming and XR applications, latency and bandwidth constraints are far more severe than in traditional streaming. V-Nova and NVIDIA demonstrated at NAB 2024 that LCEVC enhancement applied to NVENC HEVC reduced bandwidth requirements for a popular XR application by 55%, from 55 Mbps to 25 Mbps. In low-latency cloud gaming scenarios, LCEVC achieved 30% compression gains while operating within the tight encode-decode cycles that interactive applications demand. The technology also powers six-degree-of-freedom (6DoF) VR streaming, including that of V-Nova's PresenZ media format, enabling photorealistic volumetric content delivery to lightweight client devices.



Brazil's TV 3.0 is built on ATSC 3.0 physical layer technologies, and that architectural alignment has direct implications for the United States. As of late 2024, 76% of the US population was within reach of an ATSC 3.0 NextGen TV signal, with broadcasts available in nearly 70 markets. LCEVC is already included in the ATSC A/345 standard specification, providing a future-proofed path for US broadcasters to upgrade their services. Ateme's TITAN Live encoder, already deployed in over 40 US markets for ATSC 3.0 NextGen TV transmission, has already demonstrated support for LCEVC at the 2025 ATSC NextGen TV conference. By integrating LCEVC, broadcasters can reduce the bitrate required for 4K HDR content by up to 40% while simultaneously lowering encoding complexity and cost. The technology's validation in Brazil's live Olympic broadcast provides operational confidence that US deployments can reference.

The Economic Argument for Enhancement Over Replacement

The traditional codec upgrade cycle follows a predictable pattern: a new standard is ratified; hardware support lags by several years, encoding infrastructure must be rebuilt, and the industry gradually migrates over a five-to-seven-year period. During that transition, platforms must maintain parallel encoding pipelines, support multiple codec profiles, and manage the complexity of serving a fragmented device landscape.

LCEVC disrupts this cycle because it works with the codecs that platforms are already using. An operator currently deploying H.264 can gain 46% bitrate savings by adding LCEVC, without waiting for their entire device fleet to support HEVC or VVC. A platform that has already invested in HEVC infrastructure can extend its efficiency frontier by 31% without rebuilding its encoding pipeline. And when VVC eventually achieves broad device support, LCEVC can enhance it further, delivering 36–55% additional savings depending on content and configuration.

This is not a theoretical distinction. It changes how platforms plan their technology road maps. Instead of large, discontinuous infrastructure investments timed to codec transitions, operators can pursue continuous, incremental efficiency improvements that compound over time. The encoding infrastructure investment is preserved rather than stranded. The device compatibility challenge is reduced rather than replicated.

THE BIGGER PICTURE

The environmental implications deserve attention. LCEVC's ability to reduce transcoding compute requirements by 40 to 70% at equivalent quality translates directly into lower energy consumption for encoding operations.

For large-scale streaming platforms processing millions of hours annually, the aggregate power reduction is substantial. As the industry faces growing pressure to reduce its carbon footprint, codec efficiency becomes an environmental question as well as an economic one.

The Strategic Calculus

The video industry does not lack codec technologies. Missing however is the clear path from standardization to deployment that accounts for the realities of installed infrastructure, fragmented device ecosystems, and competing commercial priorities. LCEVC's value is not that it outperforms any single codec in isolation. It is that it makes every codec it touches perform better, on devices that already exist, through infrastructure that is already in place.

Brazil's TV 3.0 deployment is the proof that this approach works at national scale. The Paris 2024 Olympics broadcast demonstrated it in live production. The 50-company ecosystem at NAB 2025 demonstrated it across the full delivery chain. And the independent research, from MPEG's own verification tests to peer-reviewed academic studies, confirms that the efficiency gains are real, reproducible, and consistent across content types, codecs, and viewing conditions.

For streaming platforms, broadcasters, and pay TV operators evaluating their next infrastructure investment, the question is not whether LCEVC works. The evidence on that point is settled. The question is whether it makes strategic sense to continue pursuing quality improvements through codec replacement alone, when a standards-based enhancement technology can deliver measurable gains on the infrastructure you have already built.

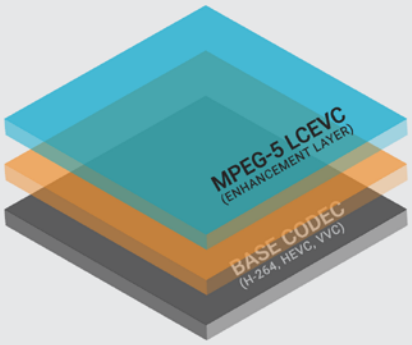
THIS ARTICLE IS PART OF THE VPU ECOSYSTEM SERIES EXPLORING HOW PURPOSE-BUILT SILICON IS RESHAPING VIDEO INFRASTRUCTURE ACROSS THE TECHNOLOGY STACK. EACH ARTICLE EXAMINES A DIFFERENT ARCHITECTURAL PERSPECTIVE THROUGH THE LENS OF ONE PARTNER WITHIN THE ECOSYSTEM.

MPEG-5 LCEVC:
Rewriting the Economics of Video Quality

THE ENHANCEMENT PARADIGM

A Codec-Agnostic Enhancement Layer

LCEVC is standardized as ISO/IEC 23094-2, designed to enhance any base codec from legacy H.264 to the latest VVC.



QUANTIFIABLE EFFICIENCY GAINS

46%

Bitrate Savings for H.264/AVC

LCEVC is standardized as ISO/IEC 23094-2, designed to enhance any base codec from legacy H.264 to the latest VVC.

52-55%

Gains for VVC 4K HDR

Peer-reviewed research shows that even the most advanced codecs see a doubling in efficiency when paired with LCEVC.

BASE CODEC	BITRATE SAVINGS	SOURCE / METHODOLOGY
HEVC	46% (UHD)	MPEG Verification (Subjective MOS)
HEVC	31% (UHD)	MPEG Verification (Subjective MOS)
VVC	52-55%	Frontiers in Signal Processing (4K HDR)
X265 (gaming)	39%	ACM Mile High Video (Subjective/VMAF)
SVT-AV1	~40%	SPIE / Intel / Meta (Convex Hull VMAF)

STRATEGIC DEPLOYMENT VECTORS

OTT & Streaming Optimization

Reduces transcoding costs and energy consumption by 70% while improving VMAF quality scores.

Pay TV & Legacy STB Longevity

Delivers 1080p HDR at SD bitrates to installed devices via firmware updates, avoiding expensive "truck rolls" or hardware swaps.

Cloud Gaming & XR

Reduced bandwidth for XR applications by 55% (from 5X Mbps to 25 Mbps) within tight low-latency windows.

NETINT VPU ECOSYSTEM

CLOSING PERSPECTIVE

The window is open. But, not for long.

The GPU monoculture is cracking.
VPU evaluation intent is within 2 points of GPU for the first time ever.

The bottleneck isn't technical. It's organizational.
Budget and team capacity block deployments 4X more than technology.

The codec transition is already underway.
AV1 will triple its footprint in 2026.
Infrastructure that can't follow will be replaced.

The Ecosystem exists to remove every excuse.
Silicon, servers, orchestration, measurement, delivery.
The full stack is here, proven, deployable.

The organizations moving now will set the standard everyone else builds toward.

Start your VPU deployment today.

NETINT.COM

NETINT TECHNOLOGIES

The market is not waiting for a better encoder.

It is 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.

That is the category NETINT is building.

VPU
ECOSYSTEM

NETINT.COM

NETINT VPU Ecosystem