More Streams, Less Power: Energy Trade-offs in VPU vs GPU Transcoding

A benchmark-driven study of performance, cost and sustainability in video infrastructure

cires21.com

September 2025

Abstract.

The surge in live and on-demand video consumption has magnified the importance of efficient and scalable transcoding infrastructure. **Graphics Processing Units (GPUs)** have been the de facto standard, but **ASIC-based Video Processing Units (VPUs)** are increasingly positioned as a disruptive alternative.

This whitepaper benchmarks the **NETINT Quadra T1U (VPU)** against the **NVIDIA RTX 4000 Ada (GPU)** in a controlled cloud environment. Tests measure **throughput, per-stream power efficiency, and perceptual quality** (VMAF) across H.264, HEVC, and AV1 codecs.

Key Results

01. **NETINT delivers 4×–6× greater energy efficiency per stream**, consuming as little as 0.4–0.7 W/stream.
02. **NVIDIA delivers higher perceptual quality at HD (1080p/720p)** but at 3–6× the power cost.
03. **At low resolutions (432p/360p), NETINT outperforms NVIDIA in both quality and efficiency**, making it ideal for ABR ladders and mobile delivery.

The findings highlight critical trade-offs for datacenters, broadcasters, and edge deployments where scaling, TCO, and sustainability targets converge.

Introduction.

Modern video distribution relies on **multi-resolution, multi-codec adaptive bitrate (ABR) workflows**. This multiplies the transcoding workload, as each channel must be processed into multiple rungs of an ABR ladder.
The challenge is to maximize **density (streams per server), visual quality, and efficiency (W/stream)** while respecting power and cooling budgets.

GPU

GPUs: versatile and mature, with broad integration in broadcast and creative pipelines. They excel at HD quality but carry high energy costs.

VPU

VPUs (ASIC-based): purpose-built for video, offering predictable scaling and remarkably low per-stream power. Their trade-off lies in narrower flexibility but superior efficiency.

This study compares both under **identical cloud conditions**, quantifying their trade-offs and architectural bottlenecks.

Methodology.

Cloud Testbed

- **Environment:** Akamai Cloud, Frankfurt region
- **VM Size: x1 Small** — identical CPU resources to minimize bias
- **Instances:**
  - NVIDIA RTX 4000 Ada (CUDA/NVENC pipeline)
  - NETINT Quadra T1U (Libxcoder pipeline)

Workflows

- **Input:** 6-minute raw YUV 4:2:0 @1080p60.
- **Test flows:**
  - 1:1 transcodes
  - 1→N ABR ladders (1080p, 720p, 576p, 432p, 360p outputs)

Measurements

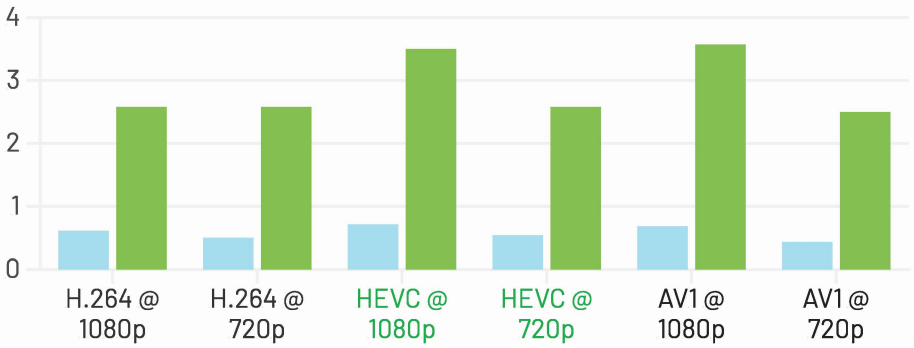
- Maximum real-time concurrent jobs
- Total and per-stream power draw
- Encoder, scaler, and GPU utilization percentages
- Perceptual quality via **VMAF (average & 5th-percentile)**

Results.

Density, Power, and Utilization

Power-Per-Stream Efficiency

Codec	Resolution	NETINT W/Stream	NVIDIA W/Stream	Efficiency Gain (NVIDIA ÷ NETINT)
H.264	1080p	0.61	2.58	4.2 ×
	720p	0.50	2.58	5.2 ×
HEVC	1080p	0.71	3.50	4.9 ×
	720p	0.54	2.58	4.8 ×
AV1	1080p	0.68	3.57	5.2 ×
	720p	0.43	2.50	5.8 ×



NETINT averages **0.4–0.7 W/stream** across all codecs; NVIDIA averages **2.5–3.6 W/stream**, worst in **AV1 at 1080p**.

Resource Utilization Insights

- **NETINT:** Encoder load ~99% at 1080p; scaler load bottlenecks ABR workflows.
- **NVIDIA:** GPU utilization 67–92%, but NVENC engines saturate early.
- **Thermals:** NETINT runs cool (~35–37°C), enabling dense rack deployments.

Transcode Job Capacity by Resolution

Across all tested codecs, both encoders demonstrated distinct scaling behavior depending on resolution:

- At 1080p, job counts converge (≈19 streams for H.264), but the GPU consumes ~4× more power for the same throughput.
- At 720p, NVIDIA scales slightly higher in stream count (24 vs. 22 for NETINT) but at a 5× energy penalty.
- At lower rungs (432p, 360p), NVIDIA reaches up to 30 streams, compared to NETINT's 20–21, yet efficiency reverses: NETINT maintains sub-0.5 W/stream, while NVIDIA requires 1.5–1.7 W/stream.

Results.

At high resolutions (1080p), NETINT achieves equal or greater capacity with much lower power draw. At mid resolutions (720p), capacities converge but efficiency gaps widen. At low resolutions (360p), NVIDIA reaches higher job counts but at an untenable energy cost, making NETINT the practical choice for sustainable scale-out deployments.

Codec	Resolution	NETINT Jobs	NETINT Power (W)	NVIDIA Jobs	NVIDIA Power (W)
H.264	1080p	19	11.67	19	49
	720p	22	11.03	24	63
	576p	20	8.64	25	55
	432p	21	7.94	28	48
	360p	20	7.64	30	45
HEVC	1080p	18	12.72	14	49
	720p	22	11.79	24	62
	576p	20	9.98	25	46
	432p	21	8.46	28	46
	360p	20	7.94	30	44
AV1	1080p	18	12.25	14	50
	720p	25	10.85	24	60
	576p	20	9.92	25	53
	432p	20	8.52	30	49
	360p	20	8.05	30	44

Energy Usage for 1,000 Streams (Annualized)

To contextualize efficiency differences at scale, consider supporting 1,000 concurrent 1080p H.264 streams continuously for one year:

Encoder	Power/Stream (W)	Total Power (W)	Annual Energy (kWh)	Relative Efficiency
NETINT Quadra T1U	0.614	~614	~5,376	Baseline
NVIDIA RTX 4000 Ada	2.579	~2,579	~22,600	~4.2× higher energy cost

Video Quality (VMAF)

Findings

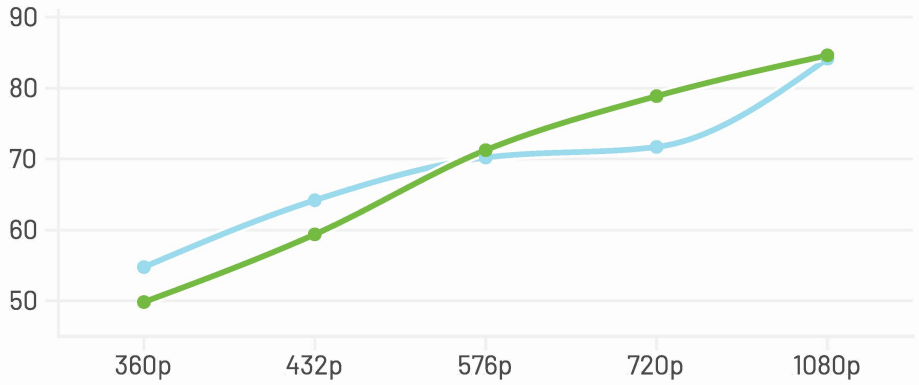
- **HD (1080p/720p):** NVIDIA dominates, up to +8 VMAF points.
- **Low res (432p/360p):** NETINT clearly outperforms by +4–6 points.
- **Codec trends:** AV1 leads at HD; NETINT's strengths grow as resolution decreases.

Results.

Detailed VMAF Score Data

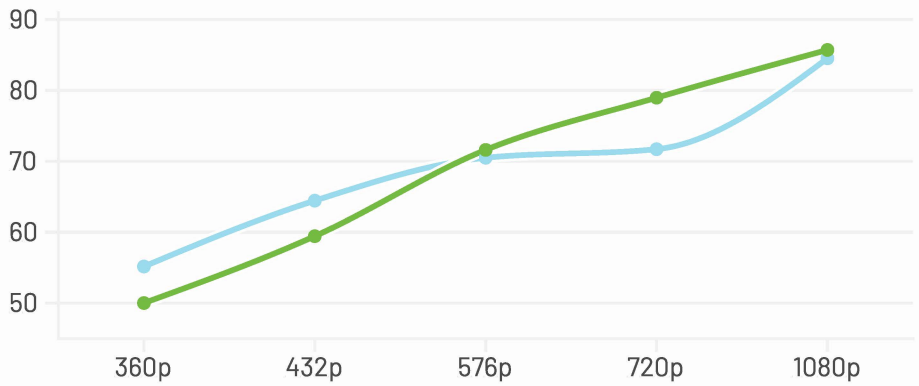
H.264

Resolution	NETINT VMAF	NVIDIA VMAF	Δ (NVIDIA - NETINT)	Winner
1080p	84.17	84.63	0.46	NVIDIA
720p	71.71	78.88	7.17	NVIDIA
576p	70.21	71.26	1.05	NVIDIA
432p	64.19	59.38	-4.81	NETINT
360p	54.76	49.83	-4.93	NETINT



HEVC

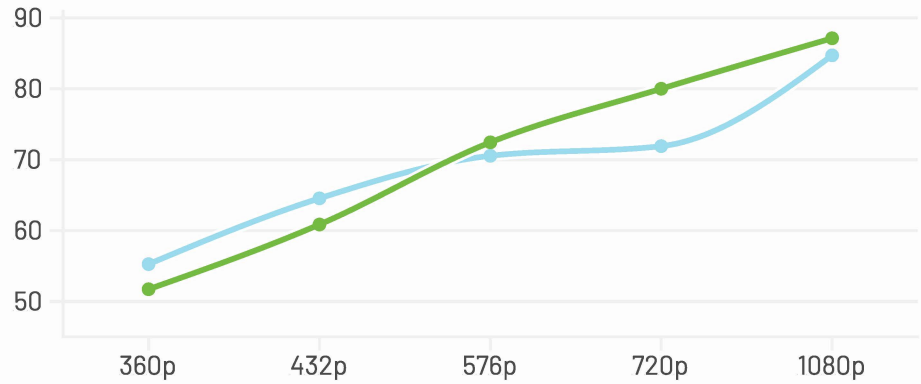
Resolution	NETINT VMAF	NVIDIA VMAF	Δ (NVIDIA - NETINT)	Winner
1080p	84.53	85.71	1.18	NVIDIA
720p	71.71	78.97	7.26	NVIDIA
576p	70.50	71.61	1.11	NVIDIA
432p	64.45	59.44	-5.01	NETINT
360p	55.16	50.01	-5.15	NETINT



Results.

AV1

Resolution	NETINT VMAF	NVIDIA VMAF	Δ (NVIDIA - NETINT)	Winner
1080p	84.72	87.13	2.41	NVIDIA
720p	71.91	80.01	8.1	NVIDIA
576p	70.55	72.45	1.9	NVIDIA
432p	64.55	60.86	-3.69	NETINT
360p	55.27	51.73	-3.54	NETINT



5th-Percentile Scores

- **NETINT:** AV1 = 55.3, HEVC = 55.2, H.264 = 54.8
- **NVIDIA:** AV1 = 51.7, HEVC = 50.0, H.264 = 49.8

NETINT delivers better worst-case quality, reducing visual degradation in constrained bandwidth scenarios.

VMAF Stability Index (Resolution Resilience)

VMAF drop from 1080p → 360p, per codec:

Codec	NETINT Drop	NVIDIA Drop
H.264	-34.9%	-41.1%
HEVC	-34.7%	-41.6%
AV1	-34.7%	-40.6%

NETINT consistently shows a ~35% drop vs NVIDIA's ~41%, proving higher resilience across ABR ladders.

Results.

Quality Retention Insights

On quality retention, the difference between encoders is not just where they score highest, but how stable their quality curves are:

- NVIDIA NVENC (Ada RTX 4000):
  - Higher average VMAF at HD.
  - Steeper quality decline below 720p, especially in AV1 and H.264.
  - Variability means more visible low-quality segments.

- NETINT Quadra T1A:
  - More gradual decline in VMAF as resolution decreases.
  - Higher 5th-percentile scores → better worst-case frames.
  - Maintains detail at 432p and 360p, critical for mobile/low-bandwidth delivery.

Conclusion: NETINT prioritizes quality consistency, while NVIDIA prioritizes HD fidelity.

Technical Discussion.

- **Bottlenecks:**
  - NVIDIA limited by NVENC engine concurrency at high resolutions.
  - NETINT limited by scaler sharing in 1→N encoding, mitigated via auxiliary 2D/Blitter scalers.

- **Operator trade-off:** At 360p, NVIDIA gains 33% more streams — but burns 6× more power.

- **Rate Control:** NETINT's capped-CRF + look-ahead produces predictable output. NVIDIA favors speed and throughput at lower resolutions.

Recommendations

- **NETINT Quadra T1U** = efficiency & resilience champion.
- **NVIDIA RTX 4000 Ada** = HD quality leader, power-hungry at scale.

Scenario	Best choice	Rationale
High-density datacenters	NETINT	4–6× energy savings per stream
Mobile / low-bandwidth streaming	NETINT	Better low-res quality retention
Live broadcast (1080p/720p)	NVIDIA	Higher perceptual quality
Edge deployments / compact racks	NETINT	NVMe form factor, low thermal footprint
AV1 adoption at scale	NETINT	Efficiency advantage increases with codec complexity

Future Outlook.

Looking forward

- **Next-gen codecs (VVC/H.266):** Both require new silicon. VPUs face power density challenges; GPUs face parallelization inefficiency.
- **On-board AI cores:** NETINT supports ROI coding & HVS+ optimizations at runtime.
- **Sustainability:** VPUs offer the most scalable path toward green broadcasting.

Final thoughts

The choice between GPU and VPU is contextual:

- **NVIDIA GPUs** deliver superior HD perceptual quality and are best suited for **premium contribution workflows** where fidelity trumps all else.
- **NETINT VPUs** offer unmatched efficiency and resilience, making them the clear choice for **scalable ABR ladders, mobile delivery, and datacenter deployments**.
- A hybrid future — VPUs for density, GPUs for flagship streams — may define the next decade of transcoding infrastructure.

At scale, the difference is not marginal. Supporting **1,000 concurrent 1080p streams for one year** requires **~5,400 kWh on NETINT** vs ~**22,600 kWh on NVIDIA** — a gap of **over 17,000 kWh**. This isn't just about power bills: it is about sustainability targets, cooling infrastructure, and the practical limits of data center density.

In practice, the industry is likely to converge on a **hybrid approach**: VPUs for efficiency and bulk capacity, GPUs reserved for premium tiers where every pixel matters. This balance maximizes both economics and viewer experience while ensuring future-proof, energy-conscious infrastructures.

Acknowledgements

This study was conducted by Cires21 in collaboration with NETINT Technologies and Akamai, leveraging reproducible cloud benchmarks in real-world workflows.
Special thanks to Dennis Mungai (Cires21), Chris Milsted (Akamai) & Mark Donnigan (NETINT)
Edited by Nacho Mileo (Cires21)
Photo cover by Mateus Durães dos Santos (Unsplash)