GPU SMITH
COMPUTE ENGINEERING

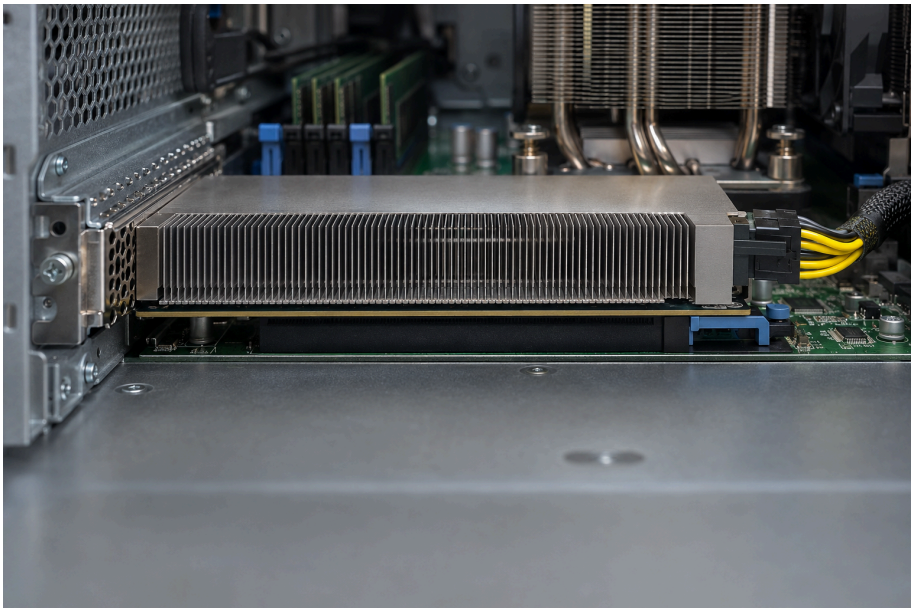
GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

Video Preprocessing Cost: NVIDIA L4 vs CPU vs MediaConvert

> Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-11-08 · video preprocessing cost · nvidia l4 transcoding · gpu video transcoding · aws mediaconvert pricing · cpu vs gpu transcoding · ffmpeg nvenc · x264 · private video dataset · ai data pipeline · self-hosted transcoding

Original article: https://gpusmith.com/articles/en/video-preprocessing-cost-l4-vs-cpu-vs-mediaconvert



In this article

1. Executive Summary
2. Introduction and Background
3. CPU Workers (libx264 on General-Purpose Instances)
4. NVIDIA L4 (Rented or Owned)
5. AWS Elemental MediaConvert
6. Feature Comparison
7. Performance and Benchmarks
8. Data Analysis and Evidence
9. Implications and Future Directions
10. Frequently Asked Questions (FAQs)
11. Conclusion

Executive Summary

This report prices one concrete job: normalizing 1,000 source-hours per month of 1080p30 H.264 footage into three H.264 single-pass renditions (1080p, 720p, 480p) within a seven-day ingest deadline, for a private AI dataset. It compares **CPU workers**, rented or owned **NVIDIA L4** capacity, and **AWS Elemental MediaConvert**. No benchmark was run for this report; throughput figures below are labeled planning inputs, and every price is a dated list price that must be re-quoted before purchase. Only MediaConvert and the rented-L4 and CPU hourly rates are priced here; AWS G6 and owned-server L4 costs are not priced, and L4 and CPU cost per source-hour appears only as a grid of assumed throughputs, so the MediaConvert figure is the only computed total.

The managed baseline is computable. MediaConvert Basic tier bills 2x normalized minutes for each HD output and 1x for SD at 30 fps or below, so three outputs from 60,000 source-minutes yield 300,000 normalized minutes. At the US East (Ohio) rates of $0.0075 for the first 100,000 ((aws.amazon.com)) and $0.0053 for the next 900,000 ((aws.amazon.com)), the service-only bill is **$1,810 per month**, or **$1.81 per source-hour**, before storage, transfers, rejects and validation. A single rented **L4** at **$0.59 per hour** ((www.runpod.io)), on a page updated November 4, 2026 ((www.runpod.io)), costs **$430.70** if left running for the 730-hour month ((aws.amazon.com)), but that bill says nothing about whether seven days is enough.

Capability gates come before price. The L4 has two NVENC engines and four NVDEC engines ((developer.nvidia.com)), while the A100 lists zero NVENC engines ((developer.nvidia.com)), so tensor throughput does not establish codec suitability. Engine counts are not sessions per second, and NVIDIA's published per-engine tables were measured on GeForce-class cards. Public vendor-run comparisons disagree: AWS reported about 73% better price/performance for a T4-based g4dn.xlarge GPU instance (not an L4) against a c6i.4xlarge CPU instance on a 4K60 batch test whose CPU and GPU presets AWS said were not exactly equivalent ((aws.amazon.com)) while warning that storage and transfer costs rise with GPU output, and a USENIX study found CPU throughput per cost higher in live transcoding but GPU higher in archive transcoding ((www.usenix.org)) ((www.usenix.org)).

The decision rule is a ratio, not a ranking. Using Google Cloud list prices as a CPU anchor, an 8-vCPU C4 instance at $0.39534 per hour ((cloud.google.com)) is beaten by an L4 priced at $0.59 per hour only if the L4 worker accepts at least **1.49 times** the source-hours per worker-hour. A bundled G2 L4 instance at $0.7068 per hour ((cloud.google.com)) needs **1.79 times**. Against MediaConvert's $1.81 per source-hour, a self-run worker at $0.59 per hour breaks even on compute at 0.33 accepted source-hours per worker-hour, so the real contest is engineering and operations cost versus a metered service. If the AI consumer only needs decoded frames, the encode step, and possibly the managed service, can be removed. The recommended path is a pilot per workload cell: pilot L4 for high-volume, deadline-bound, H.264 4:2:0 batches; retain CPU where filters or formats fall back; keep MediaConvert for bursts and small or irregular volumes.

Introduction and Background

Private video datasets (inspection footage, retail cameras, medical or industrial recordings) usually arrive in mixed containers, bitrates and frame rates. Before annotation, review or model ingestion, platform teams commonly normalize them: fix the frame rate, resize to a ladder of resolutions, keep audio and timestamps, and write a consistent codec. This is codec and data preparation, not image or video generation, event detection, or vision-encoder capacity planning. The relevant hardware question is therefore whether a dedicated media engine, a pool of CPU cores or a managed service delivers the required accepted output at the lowest complete cost.

The audience is enterprise private-video dataset engineers, computer-vision infrastructure owners and GPU procurement teams. They commit real budgets to this step, and they are often steered toward high-FLOPS accelerators by inference-centric comparisons that say nothing about codec engines.

GPU Smith is an independent private AI infrastructure engineering consultancy that advises on GPU procurement, workload sizing, networking and air-gapped deployments. It is not a cloud provider, chip vendor or transcoding service, and it is not one of the options compared here. The report applies a procurement-style discipline: define the output contract, gate by codec capability, measure the whole pipeline, then price accepted source-hours rather than peak encoder frames per second.

The worked scenario is a planning scenario, not a measured customer workload:

- **Volume:** 1,000 source-hours per month, 1080p30 H.264 input.
- **Deadline:** seven days (168 hours) from ingest to accepted output.
- **Output contract:** three H.264 single-pass renditions at 1080p, 720p and 480p, all actually needed for annotation, review or ingestion.
- **Fixed constraints:** frame sampling, audio preservation, timestamps, rotation and color handling, and downstream task quality are held constant across options.
- **Required rate:** 1,000 source-hours in 168 hours is 5.95 source-hours per wall-clock hour, about 179 source frames per second, before any failure reserve (at a 20% reserve, 7.14 times real time).

Where the AI consumer needs only decoded frames or one normalized copy, unused renditions should be removed rather than invented to justify an accelerator. Prices come from vendor pages fetched on November 8, 2026; AWS hourly G6 prices were not obtainable as page text, so they are not priced in this report.

CPU Workers (libx264 on General-Purpose Instances)

Capabilities

Software H.264 encoding with **libx264** through FFmpeg is the reference implementation most quality comparisons start from. FFmpeg's guide describes constant rate factor (CRF) as the mode for keeping quality while caring less about file size, notes that increasing CRF by 6 roughly halves the bitrate ((trac.ffmpeg.org)), and says slower presets give better compression, defined as quality per file size ((trac.ffmpeg.org)). Two-pass encoding is aimed at hitting a target file size more accurately rather than improving quality ((trac.ffmpeg.org)). FFmpeg's guide advises using the slowest preset one has patience for ((trac.ffmpeg.org)), which is exactly the lever that trades CPU time for bytes. Because rate control, preset and CRF all move cost, retained bytes and quality at once, any CPU-versus-GPU comparison must hold them fixed.

CPU capacity is sold in vCPUs, which are not uniform. On AWS, a c7i.xlarge exposes 4 vCPUs as 2 cores with 2 threads per core ((docs.aws.amazon.com)), while a c7a.xlarge exposes 4 vCPUs as 4 cores with one thread each ((docs.aws.amazon.com)). Throughput per vCPU therefore differs by family, and a measured result on one instance type does not transfer to another.

Cost Structure and Adoption Evidence

Google Cloud lists on-demand prices of $0.39534 per hour for an 8-vCPU C4 instance ((cloud.google.com)), $0.201608 for a 4-vCPU C3 ((cloud.google.com)) and $0.417616 for an 8-vCPU C2 ((cloud.google.com)). Dividing by vCPU count gives roughly $0.049 to $0.052 per vCPU-hour, which is arithmetic on list prices in the pages' default region view, not a quotation for any specific region. NVIDIA's own benchmark methodology uses x264 as the CPU reference, pairing the medium preset with a latency-tolerant NVENC configuration ((developer.download.nvidia.com)), and fixes x264 thread counts by resolution. Wowza, for its own streaming workflow, recommends keeping CPU use under 85 percent in production ((www.wowza.com)), a reminder that planned utilization is below benchmark utilization.

Strengths and Limitations

- **Format coverage:** software decode and filtering handle arbitrary pixel formats, rotation, color conversion and custom filters without moving frames between memory spaces.
- **Output size:** vendor testing summarized under Performance and Benchmarks reports smaller files from CPU encoders, which matters when files are retained for a year.
- **Licensing note:** FFmpeg states that if its optional GPL parts are used, the GPL applies to all of FFmpeg ((ffmpeg.org)), and notes it is not legal advice ((ffmpeg.org)), adding that it is not available under any other licensing terms ((ffmpeg.org)); teams building internal pipelines should have counsel review their build and distribution model.
- **Limitation:** compute cost per accepted source-hour depends entirely on measured throughput, and CPU pools must be sized for the seven-day deadline plus reserve, which may leave fleets idle outside ingest windows unless billed per use.

NVIDIA L4 (Rented or Owned)

Capabilities

NVIDIA's **support matrix** lists the L4 (Ada Lovelace) with two NVENC engines and unrestricted concurrent sessions, with H.264 4:2:0 encode supported and H.264 4:2:2 encode not supported ((developer.nvidia.com)), and four NVDEC engines ((developer.nvidia.com)). The L40S lists three NVENC engines ((developer.nvidia.com)), while the A100 lists none, as noted above; the USENIX ATC 2024 authors excluded the A100 from their video experiments for exactly that reason ((www.usenix.org)). NVIDIA's SDK page states that these engines sit separate from the CUDA cores ((developer.nvidia.com)), so encode load does not consume inference compute, but it also means that a bigger compute GPU adds nothing for transcoding.

Decode has its own gates. NVIDIA's NVDEC application note shows H.264 High10 and High422 profile decode as unsupported in the Ada column ((docs.nvidia.com)), and the FFmpeg wiki observes that hardware decoders tend not to implement anything beyond 8-bit 4:2:0 for H.264 ((trac.ffmpeg.org)), although NVDEC offers decoders for many other codecs including AV1 ((trac.ffmpeg.org)). Sources recorded as 10-bit or 4:2:2 therefore need a CPU decode fallback. With FFmpeg's generic hardware-acceleration option, unsupported streams are decoded in software automatically ((trac.ffmpeg.org)), whereas wrapper decoders do not support any fallback ((trac.ffmpeg.org)). That silent fallback is exactly what an accepted-output audit must log.

Pipeline Mechanics

NVIDIA documents flags that keep decoded frames in GPU memory for transcoding ((docs.nvidia.com)), but software filters may share no formats with hardware filters ((trac.ffmpeg.org)), so a custom CPU-side filter can reintroduce copies. A single encode session cannot exceed the performance of one NVENC engine ((docs.nvidia.com)), so a two-engine card needs several concurrent sessions; NVIDIA's FFmpeg guide suggests running several simultaneous sessions to measure aggregate throughput ((docs.nvidia.com)). The 1:N ladder case (one decode, three scaled encodes) is the closest match to this scenario and has no published L4 H.264 number in the sources reviewed.

Cost Structure and Adoption Evidence

- **Rented L4:** Runpod's page lists L4 at $0.59 per hour ((www.runpod.io)) and L40S at $1.09 per hour ((www.runpod.io)), with the FAQ describing per-second billing with no commitment for its flagship H100, A100 and H200 rentals; the pricing table offers a per-second view, but L4 billing granularity should be confirmed before purchase ((www.runpod.io)). The page's structured data lists the $0.59 rate as the Secure Cloud price for L4 and a lower $0.44 per hour for Community Cloud, so security terms and capacity must still be confirmed.
- **Bundled cloud L4:** Google lists a G2 instance with one L4, 4 vCPUs and 16 GiB at $0.706832276 per hour ((cloud.google.com)) and an 8-vCPU, 32 GiB variant at $0.853624312 ((cloud.google.com)); these machine types are excluded from sustained-use discounts ((cloud.google.com)). AWS G6 instances use L4 GPUs, and a g6.xlarge has 4 vCPUs and 16 GiB ((docs.aws.amazon.com)); the hourly G6 price was not extractable as text from AWS's dynamic pricing pages and should be taken from a dated quote.

- **Owned L4:** NVIDIA lists a 72 W maximum power for the card ((www.nvidia.com)). At 730 hours, the board alone draws 52.56 kWh, which is $7.61 per month at the August 2026 US commercial average of 14.47 cents per kWh ((www.eia.gov)), $4.71 at Texas's 8.96 cents ((www.eia.gov)) and $15.78 at California's 30.02 cents ((www.eia.gov)); EIA marks 2026 values as preliminary ((www.eia.gov)). Server, card, rack, support and refresh quotes were not obtained, so owned-server capital cost is not priced in this report.
- **Air-gapped operation:** NVIDIA's GPU Operator documentation states that a fully disconnected install requires all images to be hosted in a local image registry ((docs.nvidia.com)), an operations task to include in owned-server costs.

NVIDIA's own L4 throughput claim concerns AV1 low-latency streaming on eight cards ((www.nvidia.com)), and its September 2024 pipeline study measured HEVC on 1080p60 and 2160p60 content, finding GPU encoding two to four times cheaper than CPU x265 for specific tunes ((developer.nvidia.com)) ((developer.nvidia.com)). Neither is H.264 offline transcoding of 1080p30 footage, so neither can be read as a quote for this scenario.

Strengths and Limitations

- **Strength:** dedicated engines can take encode load off CPU cores and off the inference GPU, and low-profile, low-power cards suit dense servers.
- **Limitation:** capability gates (profile, chroma, bit depth) and the output-size tradeoff are workload specific, and the A100-class "more FLOPS" reasoning does not apply.
- **Limitation:** pay-for-capacity models bill idle hours; a continuously running card costs $430.70 whether or not the deadline is met.

AWS Elemental MediaConvert

Capabilities

MediaConvert is AWS's file-based transcoding service, billed in normalized output minutes. AWS describes each transcode as assigned a multiplier based on frame rate, resolution, codec and number of passes ((aws.amazon.com)). The Basic tier supports AVC outputs with speed-optimized or balanced single-pass encoding ((aws.amazon.com)), which matches the default contract here; multi-pass AVC and most other codecs fall outside it and carry different multipliers and the Professional tier's rates. The AWS price-list file shows the Professional tier's single-pass SD AVC entry at $0.0120 in Ohio and $0.0136 in Frankfurt, against $0.0075 and $0.0085 for Basic ((b0.p.awsstatic.com)), so tier eligibility matters before any arithmetic.

Operational limits are published, but speed is not. The account-wide concurrent-job quota, which spans all on-demand queues, is 1,000 in regions such as Ohio, and the per-queue limit there is 250 concurrent jobs (adjustable through Service Quotas); neither figure is a throughput number ((docs.aws.amazon.com)), while AWS states that job completion time varies significantly with input files and job settings ((docs.aws.amazon.com)). No per-job latency or frames-per-second figure was found, so the seven-day deadline cannot be proven from documentation and needs a pilot batch.

Cost Structure and Adoption Evidence

Tiered rates apply to total monthly normalized minutes per region and account: in US East (Ohio) the first 100,000 cost $0.0075 and the next 900,000 cost $0.0053, as shown in the Executive Summary arithmetic. AWS notes that volume discounts are calculated monthly ((aws.amazon.com)) and apply independently by region and account ((aws.amazon.com)), so a one-off project that straddles a month boundary resets the tier. Under the same tier, SD costs 1x and HD 2x at 30 fps or below ((aws.amazon.com)) ((aws.amazon.com)). Reading from S3 in the same region incurs no data-transfer charge ((aws.amazon.com)), and output encryption at rest is available through S3 server-side encryption ((docs.aws.amazon.com)), though residency and data-boundary terms for private footage still require a legal and security review.

Managed alternatives price differently. Google's **Transcoder API** charges per minute of output ((cloud.google.com)), computed for each video stream encoded ((cloud.google.com)), at $0.015 for SD ((cloud.google.com)) and $0.03 for 720p to 1080p ((cloud.google.com)). Three outputs cost **$0.075 per source-minute**, or **$4,500** for 60,000 source-minutes at list rates. Cloudflare Stream states that ingress and encoding are free ((developers.cloudflare.com)), but its model is prepaid storage at $5 per month per 1,000 minutes of capacity ((developers.cloudflare.com)) plus delivery, and Mux bills Plus-quality input at $0.03125 per minute at 1080p for the first 5,000 minutes ((www.mux.com)); both are video-delivery platforms and are shown only as context, not as dataset-normalization options.

Strengths and Limitations

- **Strength:** no capacity planning, idle cost or driver and FFmpeg build maintenance; the bill scales with output minutes.
- **Strength:** elasticity within quotas can absorb irregular volumes.
- **Limitation:** metered cost grows linearly with the number of renditions, and the service cannot be tuned beyond its presets.
- **Limitation:** data must reside in or move to AWS storage, so transfers from other locations and data-boundary terms are part of the decision.

Feature Comparison

Table 1. Capability gates by option. The table summarizes the capability gates that must be passed before cost is compared. Entries are documented capabilities, not measured performance.

Dimension	CPU workers (libx264)	Rented or owned L4 (NVENC/NVDEC)	MediaConvert Basic tier
H.264 4:2:0 8-bit output	Supported by libx264 through FFmpeg; most players need 4:2:0 H.264 ((trac.ffmpeg.org))	Encode supported; 4:2:2 encode not supported on L4	Basic tier supports AVC single-pass outputs
10-bit or 4:2:2 H.264 input	Software decode handles it; throughput must be measured	High10/High422 decode unsupported on Ada; CPU fallback needed	Check supported input list before committing; not verified here

Dimension	CPU workers (libx264)	Rented or owned L4 (NVENC/NVDEC)	MediaConvert Basic tier
Capacity unit	vCPU or core hours; threads per core vary by family	NVENC and NVDEC engines, not guaranteed sessions per second	Normalized output minutes
Pricing unit	Instance-hour, for example $0.39534 for an 8-vCPU C4	Instance-hour, for example $0.59 or $0.7068	$0.0075 then $0.0053 per normalized minute
Speed guarantee	None; measured	None; measured with several sessions	None documented
Custom filters	Unrestricted	Copies appear when software filters meet hardware frames	Preset-bound
Fallback behavior	Not applicable	Silent for generic acceleration; none for wrapper decoders	Service-managed
Data boundary	Operator-controlled	Operator-controlled if owned; provider terms if rented	AWS region and account; encryption is opt-in per output
Ops burden	Fleet scheduling and builds	Drivers, FFmpeg build and air-gap images	Job submission and quotas

The matrix shows why a cost-only ranking is incomplete. The L4 path gains from dedicated engines but adds capability and fallback checks, MediaConvert removes operations but fixes the preset and the price per rendition, and CPU keeps flexibility at the price of per-core throughput that has to be measured. Each cell of an actual workload (codec mix, bit depth, filters) can pass or fail a column independently.

Performance and Benchmarks

What the Public Evidence Does and Does Not Say

No public benchmark found in this research measures H.264 1080p30 offline transcoding to a three-rendition ladder on an L4 against CPU x264 at matched quality. What exists is vendor-run or adjacent:

- **NVIDIA methodology:** its benchmark document compares encoders using the bitrate actually generated rather than the target ((developer.download.nvidia.com)), a practice any matched-quality test should copy. NVIDIA's per-engine encode and decode tables cite measurements on an RTX 4090 for Ada ((docs.nvidia.com)), so they are not L4 numbers.
- **AWS compute study:** it compared CPU and GPU instances on 4K, 60 fps source video ((aws.amazon.com)), reported roughly 73% better price/performance for a T4-based g4dn.xlarge instance in x264 batch tests (AWS noted that the presets were not exactly equivalent between CPU and GPU and that NVENC H.264 output generally looked better), found that CPU encoders do a particularly good job of containing output file sizes ((aws.amazon.com)), cautioned that data transfer and storage cost will noticeably increase if GPUs are used ((aws.amazon.com)), and stated that the test included no specific quality measurements ((aws.amazon.com)).

- **Wowza benchmark:** captured in August 2023 from a 1080p H.264/AAC source ((www.wowza.com)), it addresses a simple RTMP to HLS streaming scenario ((www.wowza.com)), says its numbers are for guidance and reference only ((www.wowza.com)) and that other use cases can drastically change them ((www.wowza.com)).
- **USENIX ATC 2024:** the benchmark suite is built on vbench ((www.usenix.org)); the CPU server showed higher throughput per cost than A40 GPUs for live transcoding, while GPUs provided higher throughput per cost for archive transcoding (as described in the Executive Summary). The same paper notes GPUs account for a substantial portion of the traditional server's capital expenditure ((www.usenix.org)). It did not test an L4 or cloud instances.

These results do not contradict one another once the workload is specified: live, low-latency and archive-style transcodes stress different parts of the pipeline, and the hardware, codec and quality target changed between studies. They are inputs for the test design, not for the answer.

Quality Is Not Interchangeable

The FFmpeg wiki says hardware encoders typically generate output of significantly lower quality than good software encoders like x264 ((trac.ffmpeg.org)). A 2025 arXiv paper reports that NVENC provides comparable rate-distortion performance at considerably lower power, for its tested settings ((arxiv.org)), and that split-frame encoding nearly doubles throughput with a negligible penalty ((arxiv.org)). These claims differ in codec, resolution, preset and metric, so none should be applied to an unmeasured L4 H.264 configuration.

A defensible acceptance test uses several layers:

1. **Reference quality:** Netflix's **VMAF** is a perceptual video quality algorithm ((github.com)) available as an FFmpeg filter ((github.com)) alongside PSNR, SSIM and other metrics ((github.com)); the filter takes the distorted video first and the reference second ((ffmpeg.org)) and offers frame subsampling to cut QA compute ((ffmpeg.org)).
2. **Model scope:** the v1 standard model is calibrated for 1080p viewed at 3H ((github.com)), and the project warns against comparing absolute scores across native resolutions ((github.com)), so 720p and 480p renditions need their own reference handling.
3. **Structural checks:** frame counts, timestamps, rotation, color tags and audio streams, which perceptual scores do not cover.
4. **Downstream task check:** compression studies, which span pose estimation, segmentation, detection, action recognition and depth estimation ((arxiv.org)), find task-dependent effects. One found detection robust to moderate compression ((arxiv.org)) but degrading appreciably at higher levels in complex scenes ((arxiv.org)); another saw significant drops at an H.264 CRF of 40 ((arxiv.org)) with retraining recovering up to 78.4% in some cases ((arxiv.org)); and an action-recognition study reported losses of up to about 1% at CRF 30 ((arxiv.org)). Perceptual similarity is therefore not model-quality equivalence.

Measuring the Whole Pipeline

The metric that matters is accepted source-hours per worker-hour, measured over read, demux, decode, filter and resize, encode, validate and write, with all three outputs, concurrent streams, CPU and RAM headroom, storage I/O and any fallback. NVIDIA's measurement guidance warns that per-process initialization becomes

significant when transcoding many short chunks ((docs.nvidia.com)); if the dataset is made of short clips, that overhead belongs in the measurement. Pin the FFmpeg build, driver and SDK versions, since the installed driver must meet the minimum for the Video Codec SDK version FFmpeg uses ((docs.nvidia.com)), log every decode path, and keep the same preset family and rate control for both arms.

Data Analysis and Evidence

Dated Price and Tier Ledger

Table 2. Dated price and tier ledger. The table lists the inputs used below. Prices are list prices as read from vendor pages on November 8, 2026 unless noted, and regions or tiers differ.

Item	Value	Basis and date	Caveat
MediaConvert Basic, US East (Ohio)	$0.0075 first 100,000; $0.0053 next 900,000 normalized minutes	AWS MediaConvert pricing page	Frankfurt first-tier SD entry is $0.0085 ((b0.p.awsstatic.com))
Runpod L4 / L40S	$0.59 / $1.09 per hour	Page updated November 4, 2026 (Runpod pricing page)	$0.59 is the Secure Cloud rate; Community Cloud is listed lower at $0.44
Google G2 with 1 L4	$0.706832276 (4 vCPU) / $0.853624312 (8 vCPU) per hour	Google pricing page, fetched November 8, 2026	No sustained-use discount
Google C4 / C3 / C2 CPU	$0.39534 (8 vCPU) / $0.201608 (4 vCPU) / $0.417616 (8 vCPU) per hour	Google pricing pages, default region view	Region and commitment discounts differ
Google Transcoder API	$0.015 SD, $0.03 HD per output minute	Google pricing page, fetched November 8, 2026	Per stream encoded
S3 Standard, Ohio	$0.023 per GB-month, first 50 TB ((b0.p.awsstatic.com))	AWS price-list file, read November 8, 2026	JSON file, not a prose page
Internet egress (Cloud Storage)	$0.12 per GiB, first tier	Google storage pricing page	Applies only if data leaves the cloud
L4 board power	72 W maximum	NVIDIA product page	Excludes server and cooling
Commercial electricity	14.47 cents per kWh US, August 2026	EIA, preliminary	State range 8.96 to 30.02

The ledger exposes three structural facts. MediaConvert's rate steps down only after 100,000 and 1,000,000 normalized minutes, so the scenario spends most of its volume in the $0.0053 band. GPU and CPU rentals are quoted per instance hour with different CPU and memory bundles (a G2 instance bundles 4 or 8 vCPUs with the L4, while other providers allot different amounts), so the compute-per-dollar comparison is not like for like. Finally, the electricity row is small next to rental or service charges, but it scales with owned fleets and cooling that the ledger excludes.

Worked Arithmetic for the Scenario (Hypothetical Example)

All figures below are conditional on the stated contract and are not measured results.

- **MediaConvert, three outputs, Basic tier:** 60,000 source-minutes times (2 + 2 + 1) is 300,000 normalized minutes; 100,000 times $0.0075 plus 200,000 times $0.0053 is $750 plus $1,060, or $1,810, which is $1.81 per source-hour. A 5% re-encode rate would raise service cost to about $1,900.
- **MediaConvert, fewer outputs:** 1080p plus 720p (4x) is 240,000 normalized minutes and $1,492; 720p alone (2x) is 120,000 and $856; 480p alone (1x) is 60,000 and $450.
- **Google Transcoder API:** $0.075 per source-minute across three outputs is $4,500 at list prices, and $3,600 for the two HD outputs only.
- **Rented L4 worker schedule:** required rate is **5.95 source-hours** per wall-hour. Lower bound on workers is ceil(1,000 divided by (168 times measured accepted source-hours per worker-hour)). At assumed 2, 4 and 8 accepted source-hours per worker-hour, that is 3, 2 and 1 workers, consuming 500, 250 and 125 worker-hours, or $295, $147.50 and $73.75 at $0.59 per hour if billed only while working. Left running for the month, each card costs $430.70.
- **Retained bytes:** at constant total ladder bitrate of 9 Mbps (a planning assumption, not a recommendation), 1,000 hours of output is 3.6 million seconds times 9 million bits divided by 8, or 4.05 TB, costing about $93 per month at the Ohio S3 Standard first-tier rate and about $1,118 if the batch is retained for twelve months. If a GPU encode were 20% larger at equal measured quality, retention would add about $224 over twelve months per batch; this is the effect AWS cautioned about in its compute study. Moving the 4.05 TB (about 3,772 GiB) out of a cloud, where Google defines general network usage as data moving from a bucket to the Internet ((cloud.google.com)), at Cloud Storage's first-tier internet egress rate of $0.12 per GiB ((cloud.google.com)) would cost about $453, a cost that applies only when outputs leave the provider.

Planning Grid: Compute Cost per Accepted Source-Hour

Table 3. Compute cost per accepted source-hour at assumed throughputs. The table divides the hourly price by an assumed accepted throughput. Each column uses the same assumed throughput for all worker types only to expose the price per unit of work; real throughput differs by worker, and MediaConvert's price is flat regardless of speed.

Accepted source-hours per worker-hour (assumed)	8-vCPU C4, $0.39534/h	Runpod L4, $0.59/h	Google G2 L4, $0.7068/h	MediaConvert Basic, three outputs
0.5	$0.79	$1.18	$1.41	$1.81
1	$0.40	$0.59	$0.71	$1.81
2	$0.20	$0.30	$0.35	$1.81
4	$0.10	$0.15	$0.18	$1.81
8	$0.05	$0.07	$0.09	$1.81

Reading the grid correctly requires two cautions. First, the rows are not predictions: the same row applies to different worker types only if their measured throughputs are equal, which is the question under test. Second, compute is not the full bill. Add validation compute, reject and re-encode rate, storage growth, transfers, engineering and support, and for owned servers keep marginal cost (power, maintenance) separate from fully allocated cost (capital, space, staff). The break-even ratios follow directly: the Runpod L4 beats the 8-vCPU C4 on compute only above **1.49 times** the C4's accepted throughput (0.59 divided by 0.39534), the Google G2 4-vCPU L4 above 1.79 times, and the 8-vCPU G2 L4 above 2.16 times. Against MediaConvert, a $0.59 worker breaks even at **0.33** accepted source-hours per worker-hour; at about $1,810 per month the managed premium over a well-utilized worker is roughly $1,200 to $1,700 per month for assumed throughputs of 1 to 8 accepted source-hours per worker-hour, which is the budget available to pay for engineering, monitoring and risk.

Decode-Only Pipelines Are a Separate Case

If the AI consumer reads frames directly, encoding can vanish from the budget. NVIDIA positions NVDEC together with NVENC for data curation workflows that remove decode bottlenecks ((developer.nvidia.com)), DALI's video reader decodes with libavcodec or NVDEC ((docs.nvidia.com)), Decord is good at handling random access patterns common in neural-network training ((github.com)), and TorchCodec is a PyTorch-native library for decoding and encoding media ((github.com)). Some libraries need care: Decord's PyPI builds are CPU-only ((github.com)). A card without NVENC but with NVDEC engines, such as the A100 (zero NVENC engines in the support matrix), can still serve a decode-only role, so the question becomes decode throughput and frame copy cost, not rendition cost.

Implications and Future Directions

Procurement. The relevant unit is cost per accepted source-hour across the complete output contract, so purchase orders should specify a pilot batch with measured throughput, fallback logs and quality acceptance before capacity is committed. A GPU chosen on inference FLOPS may lack encoders entirely, as the A100 row in the support matrix shows, and an L4 chosen for encoders may idle during the other 23 days of a month.

Break-even surfaces. The grid reduces to inputs the buyer owns: billed hours, retained bytes, CPU fallback share, re-encode rate and engineering cost. High-volume, recurring, H.264 4:2:0 batches favor dedicated capacity if the pilot confirms a throughput multiple above the ratios computed here. Irregular volumes, mixed bit depths or short projects favor the managed service, because idle capacity and fallback consume the advantage. Costs by cell are additive: a monthly batch with a heavy CPU fallback share pays both for the card and for the CPU pool.

Data boundary. Geographic prices in this report are illustrative distinct options, not residency-equivalent deployments. For private footage, the choice among rented capacity, a managed service and an owned server includes security terms, key management and region constraints that no price table settles. Owned servers also bring local registries and image management for disconnected sites.

Newer hardware. NVIDIA's documentation lists newer-generation codec capabilities, so any 4:2:2 or high-bit-depth requirement should be re-evaluated against the exact SKU rather than assumed from the L4 row; NVIDIA states that Blackwell introduces 4:2:2 progressive and interlaced encode support in hardware

((developer.nvidia.com)). Codec mix also changes the managed bill, because the multiplier depends on codec and number of passes ((aws.amazon.com)).

Research agenda. A reproducible workbook that records the FFmpeg build, driver, input set, matched quality target, per-stage timings, fallback share and accepted output size would let different organizations compare like with like. This report does not provide that benchmark; it lays out the structure and the price inputs.

Frequently Asked Questions (FAQs)

How much does AWS Elemental MediaConvert cost per minute?

It depends on the tier, region and output (see the MediaConvert section and Executive Summary for rates and sources). Basic tier US East (Ohio) lists $0.0075 per normalized minute for the first 100,000, and each HD output counts as 2 normalized minutes per output minute while SD counts as 1 at 30 fps or below. The scenario's three outputs average $0.0302 per source-minute at blended Basic tier rates.

What does NVIDIA L4 transcoding cost?

Rented L4 capacity is listed at $0.59 per hour on Runpod and $0.7068 per hour for a G2 instance with one L4 on Google Cloud. Cost per source-hour is that price divided by measured accepted source-hours per worker-hour, so across the two L4 columns of the grid it ranges from $0.07 to $1.41, and no measured L4 H.264 figure exists in the sources reviewed.

Is GPU or CPU video transcoding cheaper per hour?

Neither wins in general. The GPU is cheaper per accepted source-hour only if its throughput multiple exceeds its price ratio, for example 1.49 times against an 8-vCPU C4. Vendor-run studies point in different directions by workload.

Is self-hosted transcoding cheaper than MediaConvert?

On compute alone, a worker at $0.59 per hour breaks even with $1.81 per source-hour at 0.33 accepted source-hours per worker-hour, so self-hosting wins on raw compute only if measured throughput reaches that level, which this report did not test. The comparison turns on utilization, operations, validation and risk.

Is FFmpeg NVENC on the L4 faster than CPU x264?

Engine counts are not throughput. NVIDIA's own per-engine tables use other hardware and one session cannot exceed one engine. The only credible answer is a pinned-build test with matched quality.

What about private or on-premises video processing?

An owned L4 is a 72 W card; disconnected sites need local registries. Server and support prices were not obtained, so record dated quotes.

Is encoding needed at all for AI preprocessing?

Not always. Decode-only pipelines can use NVDEC through libraries such as DALI, and unused renditions should be removed from the contract.

Conclusion

For a planning scenario of 1,000 source-hours per month and three H.264 renditions, MediaConvert's Basic tier yields a service-only list price of $1,810 per month, a rented L4 costs $0.59 per hour only while it is working, and CPU capacity is priced per vCPU hour with throughput that has to be measured. None of these supports a universal ranking. The engine counts of an L4 are capabilities, not throughput, the quality contract changes what "equal output" means, and retained bytes can shift the economics long after the compute bill is paid.

The practical recommendation is to decide by workload cell. Pilot an L4 path where inputs are H.264 4:2:0 8-bit, volumes recur, the deadline binds and the pilot shows a throughput multiple above the break-even ratio. Retain CPU workers where bit depth, chroma or custom filters force fallback. Use the managed service for bursts, small projects and teams without capacity to operate drivers and builds. Where the AI consumer needs only decoded frames, price the decode-only path first, because it may remove the need for both a transcoding service and an encoder.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://aws.amazon.com/mediaconvert/pricing
2. https://www.runpod.io/pricing
3. https://aws.amazon.com/calculator/calculator-assumptions
4. https://developer.nvidia.com/video-encode-decode-support-matrix
5. https://aws.amazon.com/blogs/compute/optimizing-video-encoding-with-ffmpeg-using-nvidia-gpu-based-amazon-ec2-instances
6. https://www.usenix.org/system/files/atc24-zhang-li-prototyping.pdf
7. https://cloud.google.com/products/compute/pricing/general-purpose
8. https://cloud.google.com/products/compute/pricing/accelerator-optimized
9. https://trac.ffmpeg.org/wiki/Encode/H.264
10. https://docs.aws.amazon.com/ec2/latest/instancetypes/co.html
11. https://cloud.google.com/products/compute/pricing/compute-optimized
12. https://developer.download.nvidia.com/designworks/video-codec-sdk/Video-Benchmark-Ada-July-2023.pdf
13. https://www.wowza.com/docs/wowza-transcoder-performance-benchmark
14. https://ffmpeg.org/legal.html
15. https://developer.nvidia.com/video-codec-sdk
16. https://docs.nvidia.com/video-technologies/video-codec-sdk/13.0/nvdec-application-note/index.html

17. https://trac.ffmpeg.org/wiki/HWAccelIntro
18. https://docs.nvidia.com/video-technologies/video-codec-sdk/13.0/ffmpeg-with-nvidia-gpu/index.html
19. https://docs.nvidia.com/video-technologies/video-codec-sdk/13.0/nvenc-application-note/index.html
20. https://docs.aws.amazon.com/ec2/latest/instancetypes/ac.html
21. https://www.nvidia.com/en-us/data-center/l4
22. https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=epmt_5_6_a
23. https://docs.nvidia.com/datacenter/cloud-native/gpu-operator/latest/install-gpu-operator-air-gapped.html
24. https://developer.nvidia.com/blog/enabling-customizable-gpu-accelerated-video-transcoding-pipelines
25. https://aws.amazon.com/blogs/media/optimize-costs-with-tiered-pricing-in-aws-elemental-mediaconvert
26. https://b0.p.awsstatic.com/pricing/2.0/meteredUnitMaps/elementalmediaconvert/USD/current/elementalmediaconvert.json
27. https://docs.aws.amazon.com/general/latest/gr/mediaconvert.html
28. https://docs.aws.amazon.com/mediaconvert/latest/ug/working-with-queues.html
29. https://aws.amazon.com/s3/pricing
30. https://docs.aws.amazon.com/mediaconvert/latest/ug/implementing-server-side-encryption.html
31. https://cloud.google.com/transcoder/pricing
32. https://developers.cloudflare.com/stream/pricing
33. https://www.mux.com/docs/pricing/video
34. https://arxiv.org/abs/2511.18687
35. https://github.com/Netflix/vmaf
36. https://ffmpeg.org/ffmpeg-filters.html
37. https://github.com/Netflix/vmaf/blob/master/resource/doc/models_v1.md
38. https://github.com/Netflix/vmaf/blob/master/resource/doc/faq.md
39. https://arxiv.org/abs/2007.14314
40. https://arxiv.org/abs/2211.05805
41. https://arxiv.org/abs/2204.09166
42. https://b0.p.awsstatic.com/pricing/2.0/meteredUnitMaps/s3/USD/current/s3.json
43. https://cloud.google.com/storage/pricing
44. https://docs.nvidia.com/deeplearning/dali/main-user-guide/docs/operations/nvidia.dali.fn.experimental.readers.video.html
45. https://github.com/dmlc/decord
46. https://github.com/meta-pytorch/torchcodec

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the hardware library, GPU references, networking references and vendor research. Contact GPU Smith with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: https://gpusmith.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.