Cloud vs. Local Inference Economics: Mac mini, DGX Spark, and BeastMode

Mission Control committee review comparing current published cloud accelerator pricing against the locally operated Yeti Claw fleet: the Mac mini text surface, the DGX Spark, and the two CPU-backed BeastMode inference lanes. The analysis is anchored to live benchmark data already captured on this fleet, the SemiAnalysis/Nebius study provided by the user, and current published pricing from Apple, NVIDIA, AWS, Nebius, Runpod, and the U.S. Energy Information Administration.

Published May 07, 2026 · Method: 36-month amortization, California power price assumption, representative single-request token sampling, and benchmark throughput data already published to Mission Control.

Executive Summary

The Mac mini is the strongest small-scale economics play in the current fleet. Using today's Apple refurbished market anchor and a conservative max-power assumption, its fully loaded local cost is about **$0.110/h** at 24x7 utilization and **$0.307/h** at a standard 40-hour workweek. That beats every cloud GPU rate in this study, including the budget L4 baseline.

DGX Spark is a utilization-sensitive box. It looks expensive if it sits idle, but at real utilization it still undercuts the published enterprise-cloud GPU rates we checked. At 24x7 amortization it lands around **$0.254/h**; at a 40-hour workweek it rises to **$0.829/h**. That is still below Nebius L40S and far below H200-class pricing, but no longer below the cheapest L4 rental baseline.

The BeastMode lanes are best understood as spare-capacity text workers, not GPU replacements. Their CPU-only throughput is much lower, but because they ride on shared, already-owned ESXi hosts, their incremental capital burden is small and their breakeven against equivalent CPU-cloud pricing is measured in weeks, not years.

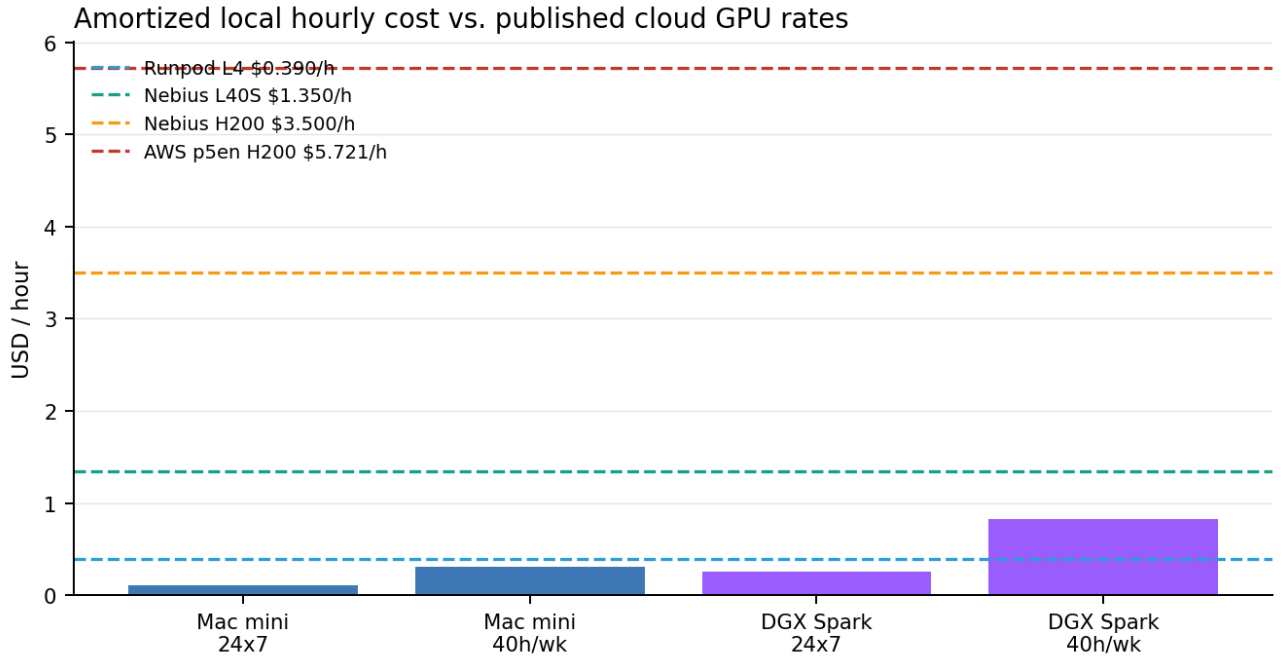
Study Context

The study you provided argues that raw $/GPU-hour is not enough. For its inference-endpoint scenario, it concludes the TCO ratio is **1.00x Nebius**, **2.13x AWS**, and **1.04x silver-tier**, even after large AWS discount assumptions. The main drivers called out are GPU instance price, support, storage, and setup overhead. Our local-fleet analysis agrees with the spirit of that result: utilization, support overhead, and surrounding infrastructure matter as much as the headline accelerator sticker price.

Study scenario	Nebius	AWS	Silver-tier	Interpretation
Large LLM pretrain	1.00x	1.09x	1.08x	AWS premium driven mainly by support and setup.
Multimodal RL research	1.00x	1.43x	1.08x	GPU instance, storage, support, and setup widen the gap.
Inference endpoints	1.00x	2.13x	1.04x	Inference TCO spreads sharply once support and surrounding ops are i

Local vs. Cloud Hourly Cost

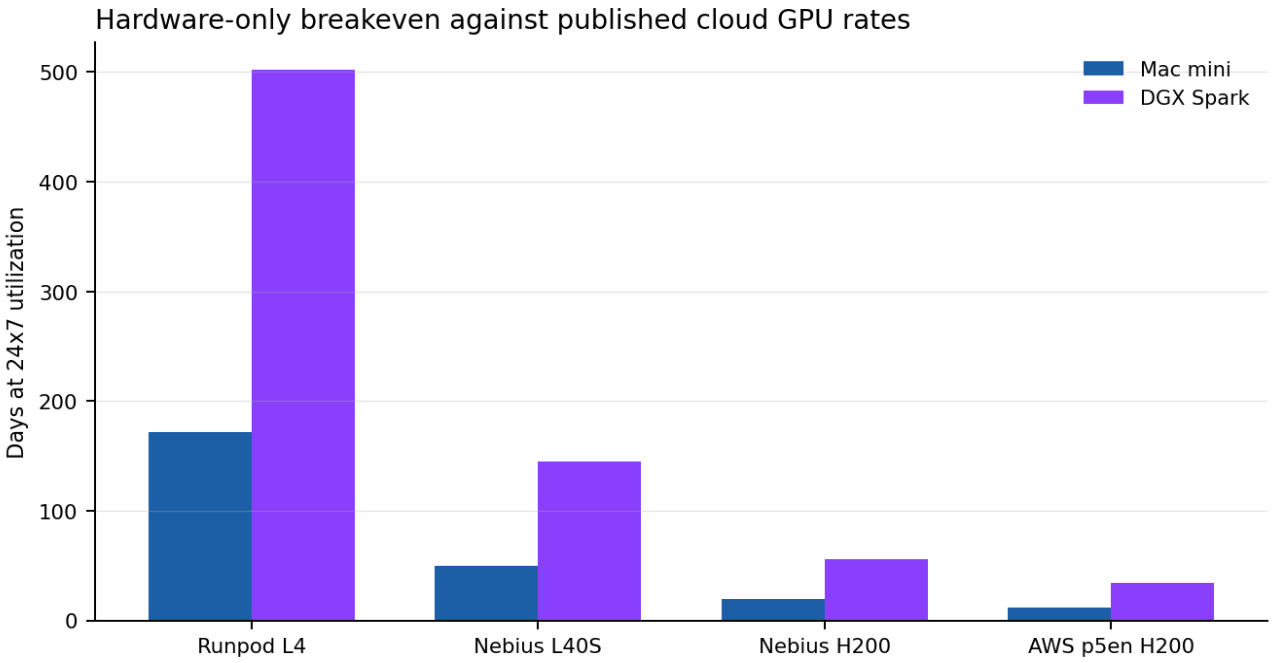
The chart below compares fully loaded local hourly cost for the Mac mini and DGX Spark against current published cloud pricing. Local costs include hardware amortization and a conservative electricity model. Mac mini uses Apple's maximum continuous power. DGX Spark uses the full 240W PSU as an intentionally conservative upper bound.



Platform	Capital anchor	24x7 local $/h	40h/week local $/h	Representative lane
Mac mini	$1,609	0.110	0.307	Interactive text inference
DGX Spark	$4,699	0.254	0.829	Text + image inference / experimentation
Runpod L4	n/a	$0.390	$0.390	Budget cloud GPU baseline
Nebius L40S	n/a	$1.350	$1.350	Enterprise mid-range GPU baseline
Nebius H200	n/a	$3.500	$3.500	Premium inference GPU baseline
AWS p5en H200	n/a	$5.721	$5.721	Hyperscaler premium GPU baseline

Breakeven Window

Breakeven here is intentionally simple: hardware sticker price divided by the published cloud hourly rate. This is the cleanest way to answer “how long until the box pays for itself?” before adding support labor or storage.

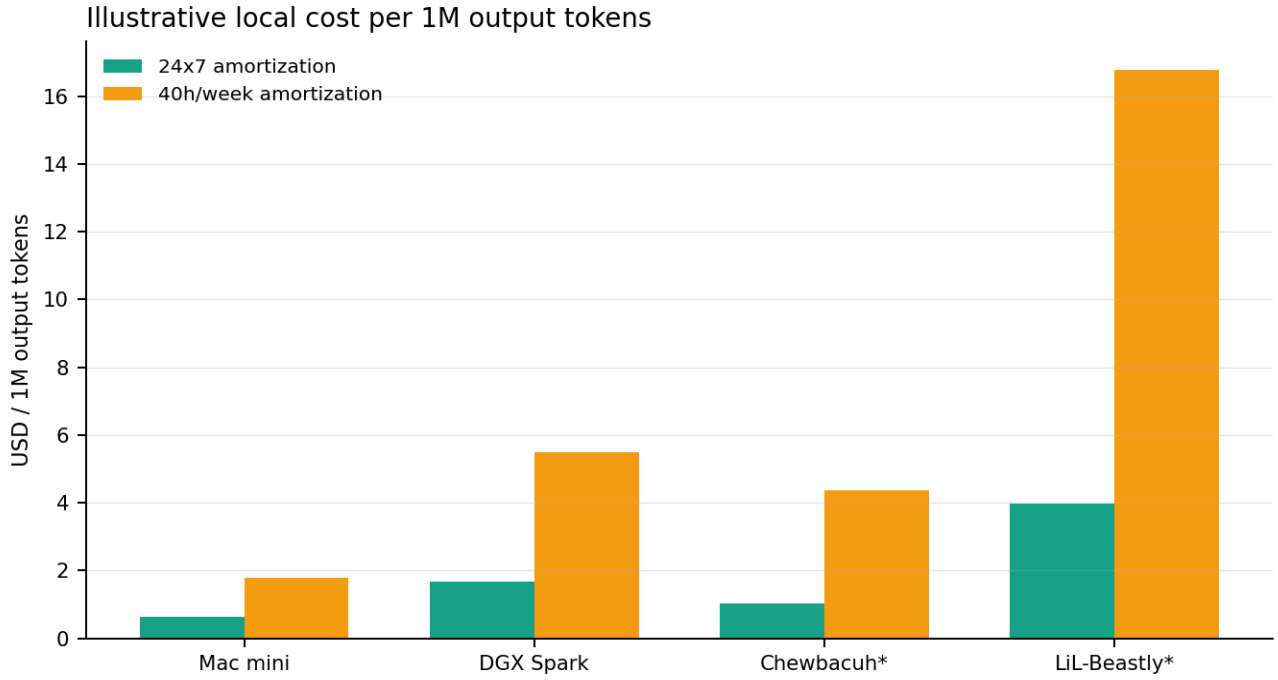


Platform	Runpod L4	Nebius L40S	Nebius H200	AWS p5en H200
Mac mini	4125.64h / 171.90d	1191.85h / 49.66d	459.71h / 19.15d	281.24h / 11.72d
DGX Spark	12048.72h / 502.03d	3480.74h / 145.03d	1342.57h / 55.94d	821.36h / 34.22d

Important read: the Mac mini pays back quickly even against cheap L4 rental if you keep it busy. DGX Spark needs meaningfully more hours to beat bargain cloud pricing, but it catches up fast against H200-class enterprise rentals.

Token Economics

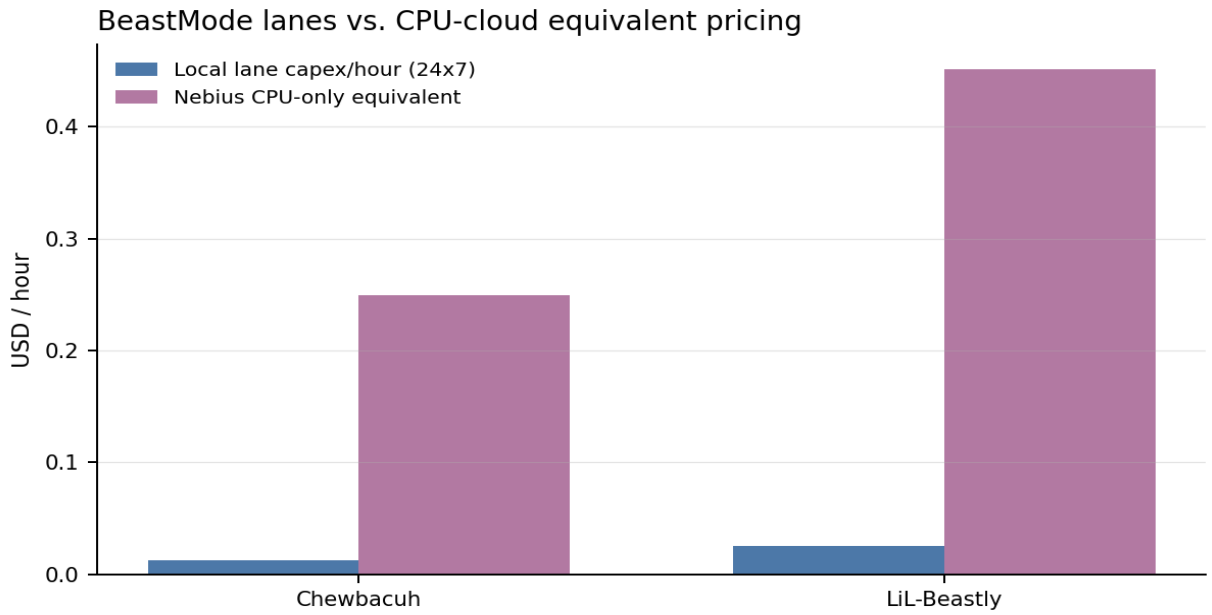
To normalize very different boxes, we sampled a representative local request on each text lane and converted output-token throughput into an illustrative cost per 1M output tokens. For the Mac mini and DGX Spark this includes power plus capital. For the BeastMode lanes it is a lower-bound capital-only view because we do not have trustworthy per-host power telemetry from ESXi in the management path used for this report.



Lane	Representative model	Output tok/s	24x7 $/1M tok	40h/week $/1M tok	Note
Mac mini	qwen2.5:7b	47.77	0.64	1.78	Best small-box economics in the fleet.
DGX Spark	qwen3:8b	41.97	1.68	5.48	Wins when shared across text and image duty cycles.
Chewbacuh*	qwen3:8b	3.32	1.04	4.37	Lower-bound, capital only.
LiL-Beastly*	qwen3:14b	1.78	3.98	16.78	Lower-bound, capital only.

BeastMode Appendix

The ESXi-backed BeastMode lanes are not GPU boxes, so the right cloud comparison is CPU-only pricing. We used Nebius non-GPU AMD EPYC pricing as a clean published baseline for equivalent vCPU and RAM footprints.



Lane	VM shape	Shared-host capital share	Nebius CPU-cloud equivalent	Breakeven
Chewbacuh	8 vCPU / 48 GiB	$326	$0.250/h	1304.75h / 54.36d
LiL-Beastly	12 vCPU / 96 GiB	$670	$0.451/h	1484.60h / 61.86d

Recommendation: keep BeastMode positioned as a queue-backed text section for users who want model variety, not as the frontline premium chat surface. For premium interactive economics, the Mac mini remains the better fit. For premium multimodal and image work, Spark remains the strategic box.

Source Desk

- **SemiAnalysis / Nebius study PDF used for TCO context**: file:///Users/nation/Downloads/nebius-semianalysis-real-cost-of-gpu-clusters.pdf — Local study provided by user; pricing snapshot in the study is February 2026.
- **Apple Certified Refurbished Mac mini (M4 Pro / 48GB / 512GB / 10GbE)**: https://www.apple.com/shop/product/g1kzmll/a/Refurbished-Mac-mini-Apple-M4-Pro-Chip-with-12-Core-CPU-and-16-Core-GPU-10Gb-Ethernet- — Used as a current-market anchor for the Mac mini capital cost.
- **Apple Mac mini technical specifications**: https://www.apple.com/mac-mini/specs/ — Used for M4 Pro architecture details and maximum continuous power.
- **NVIDIA DGX Spark marketplace listing**: https://marketplace.nvidia.com/en-us/enterprise/personal-ai-supercomputers/dgx-spark/ — Used for current published DGX Spark price.
- **NVIDIA DGX Spark hardware overview**: https://docs.nvidia.com/dgx/dgx-spark/hardware.html — Used for DGX Spark system specs and power envelope.
- **Runpod L4 cloud pricing**: https://www.runpod.io/gpu-models/l4 — Used for budget L4 cloud baseline.

- **Nebius AI Cloud compute pricing**: https://docs.nebius.com/compute/resources/pricing — Used for L40S, H200, and CPU-only cloud baselines.
- **AWS EC2 Capacity Blocks for ML pricing**: https://aws.amazon.com/ec2/capacityblocks/pricing/ — Used for p5en H200 per-accelerator pricing.
- **U.S. EIA electricity price table, California YTD through Feb 2026**: https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=epmt_5_06_b — Used for local power-cost assumption of $0.3148/kWh.
- **Newegg Dell R730xd market listing**: https://www.newegg.com/dell-poweredge-r730xd-rack/p/2NS-0008-703C5 — Used as the vm1/chewbacuh host replacement-cost anchor.
- **TechMikeNY Dell R730xd market listing**: https://techmikeny.com/products/dell-poweredge-r730xd-server-24-bay-2-20ghz-40-core-512gb-ram-26x-caddies — Used as the vm2/LiL-Beastly host replacement-cost anchor.