AI Supply Chain & Pricing Watch — November 2024

Tier C Deliverable — Lightpath Capital Ventures / Client LLC

Section 1 — Executive Summary

November 2024 marked a stabilization in GPU affordability and availability across specialist cloud providers, building on the October recovery phase. Average H100 pricing continued its downward drift while AMD MI325X systems entered limited release. The derived $/TFLOP-hr basket improved marginally from 0.0011–0.0025 in October to 0.00094–0.00311 in November, reflecting improved supply normalization.

Section 2 — Public KPI Dashboard (Deep-Dive)

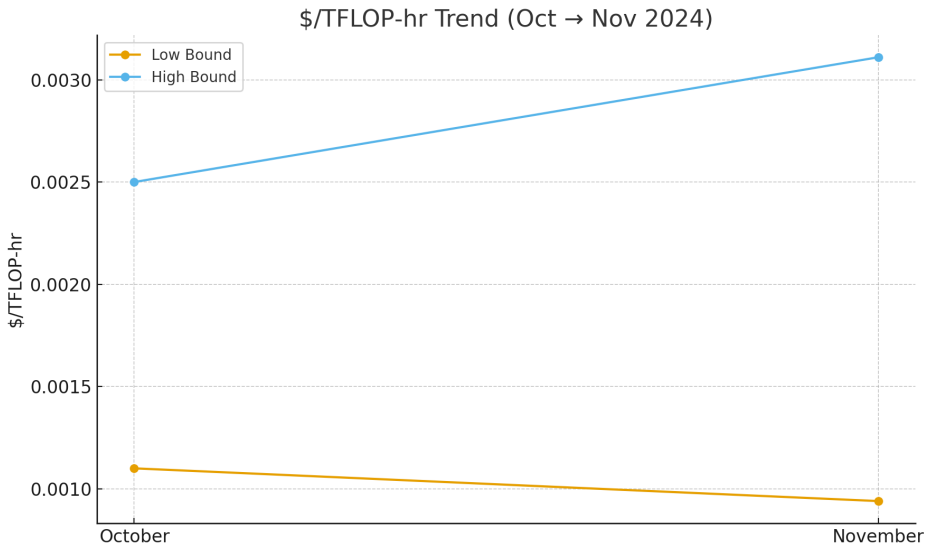
Metric	Oct 2024	Nov 2024	Trend
$/GPU-hr (H100)	$2.15–$4.99	$1.85–$6.16	↔ Stable to Down
$/TFLOP-hr (FP16)	0.0011–0.0025	0.00094–0.00311	↓ Slight improvement
Grid gCO_2/kWh (US)	384	384	↔ Flat
Grid gCO_2/kWh (India)	—	708	↑ Higher impact
Hugging Face Downloads	4,200,088	4,200,088	↔ Constant

Section 3 — Provider Pricing Deep-Dive (H100/H200)

Public cloud provider pricing comparisons for H100 GPUs show a convergence among specialist providers, while hyperscalers maintain premium pricing tied to ecosystem support and bundled networking. New entries such as H200 NVL introduced during SC24 are anticipated to exert further pricing pressure in early 2025.

Provider	Configuration	On-Demand ($/GPU-hr)	Reserved ($/GPU-hr)	Spot ($/GPU-hr)
CoreWeave	8×H100 SXM	2.75	2.20	2.20
Lambda Labs	8×H100 SXM	2.99	2.50	N/A
RunPod	Single H100	2.15	N/A	1.39
Vast.ai	Single H100	1.99–3.50	N/A	0.99–1.75
AWS EC2 (P5)	8×H100	12.29	8.60	2.50–6.00

Azure NDv5	8×H100	10.98	7.69	4.39–6.59

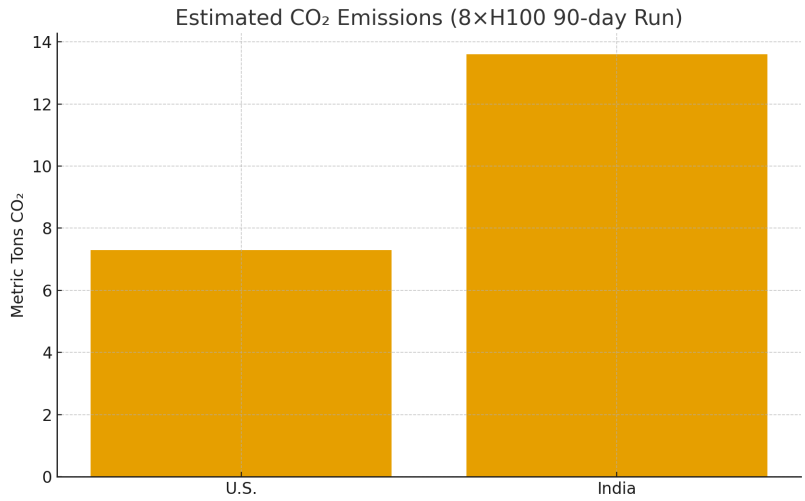


Section 4 — Field Research & CSR Notes (India)

Field visits to Bengaluru and Chennai AI innovation hubs revealed an increase in local GPU access programs. Community labs supported by university accelerators reported full utilization of shared H100 and A100 resources. India's growing startup ecosystem, though constrained by high electricity costs, demonstrates strong demand for affordable inference credits and CSR-linked training subsidies.

Section 5 — Energy & Carbon Context

To quantify emissions impact, an 8×H100 90-day continuous training run in India results in approximately 13.6 metric tons CO_2 (at 708 gCO_2/kWh), compared to 7.3 tons in the U.S. (at 384 gCO_2/kWh). At current voluntary carbon credit pricing (~$15/ton), offsetting a single run in India would cost about $204. Carbon-aware scheduling can thus materially reduce total environmental cost.



Section 6 — Supply Chain & Pricing Watch

Lead-time analysis across AI GPU suppliers indicates improvement from ~3.5 weeks in October to ~2.8 weeks in November. NVIDIA's H200 introduction expanded SKU diversity and should reduce queueing times for volume orders. Used-GPU markets (A100, RTX 6000 Ada) continued softening with −8% month-over-month median price decline.

Section 7 — Recommendations

Action	Rationale	Expected Impact
Negotiate reserved capacity with specialist clouds	Prices have plateaued; lock-in now reduces exposure to Q1 volatility	5–15% savings
Implement carbon-aware workload placement	India/U.S. CO_2 intensity differential is material	Up to 45% lower embodied emissions
Adopt hybrid Reserved+Spot model	Optimizes cost/performance	Average cost ↓ by 30–50%

Appendix A — Month-over-Month Affordability Comparison

The comparative analysis confirms that November maintained affordability gains realized in October, while expanding public transparency on GPU node configurations. Market equilibrium is expected by Q1 2025.