

ILLUSTRATIVE DRAFT GPU DECISION BRIEF | PROVISIONAL - OWNER CONFIRMATION PENDING

AWS 8 x H100 Inference Renewal Review

Rendered from a provisional fictional sample Decision Record using captured product evidence.

PREPARED FOR Northstar AI (fictional) Run: August 28, 2026		DECISION OWNER Owner pending confirmation Evidence through: August 24, 2026	
PLANNING HORIZON Next commitment decision within 60 days Next review: October 27, 2026		AUTHORITY TIER Provisional / self-service sample Provisional - owner confirmation pending	
PROVISIONAL POSTURE Review the current AWS 8 x H100 path before the next inference commitment. Keep AWS as the published baseline while validating alternatives before the next commitment.		DECISION BASIS The AWS configuration and published price are traceable for the stated workload and operating scope. No lower exact-market, technically comparable, orderable alternative has been confirmed. The present evidence supports a verification cycle, not a provider switch.	
WORKLOAD Inference, 24/7	QUANTITY AND GPU 8 x NVIDIA H100 on one instance	MINIMUM SHAPE 176 vCPU and 960 GiB memory	
GEOGRAPHY Northern Virginia operating scope	PURCHASE BASIS On-demand published catalog price	PLANNING INPUTS 730 h/GPU-month; 12% modeled idle time	
CURRENT PATH AND ALTERNATIVES CURRENT PATH RETAIN AS BASELINE AWS p5.48xlarge, N. Virginia; published on-demand, not buyer paid rate. BROADER-MARKET SIGNAL HELD FOR VERIFICATION \$3.90/GPU-hour global signal; market, terms, capacity, fees, and fit unconfirmed. IMMEDIATE PROVIDER SWITCH NOT SUPPORTABLE No comparable offer, orderability, capacity, quote, performance, or migration case.		DECISION ECONOMICS AWS PUBLISHED BASELINE ACCEPTED PUBLISHED FACT \$6.88 / GPU-hour \$55.04/instance-hour; \$482,150.40/year (8 x 730 h x 12). Catalog, not paid rate. MODELED IDLE EXPOSURE EXPLICIT MODEL \$57,858.05 / year 12% of AWS basis; planning math, not telemetry or verified savings. EXTERNAL-MARKET SIGNAL UNQUALIFIED - HELD \$3.90 / GPU-hour Global signal only; not annualized or used while exact-market review is open.	

Evidence, risks, and verification

DECISION POSTURE

When custom pricing is not available, favor public-market paths that balance time to market, workload performance, initial deployment cost, and ongoing operating efficiency. Forecast how the workload's size and scope may evolve before selecting a contract term. If a next-generation GPU platform is expected within the planning horizon, a long-term commitment may not provide the best cost or flexibility.

CONFIRMED FACTS

- 8 x H100 inference in Northern Virginia; AWS p5.48xlarge at \$55.04/instance-hour (\$6.88/GPU-hour). Evidence through August 24; some rows excluded.

OPEN CONFIRMATIONS

- Buyer paid rate and term; \$3.90 exact-market mapping; orderability, capacity, quote, fees, utilization, and performance.

UNSUPPORTED IMPLICATIONS

- No verified savings, actionable gap, availability, orderability, capacity, workload fit, provider selection, Record seal, or monitoring.

RISKS

Commercial basis risk: AWS catalog may differ from the buyer's paid or proposed rate.
Comparability risk: \$3.90 remains outside exact-market comparison until geography and terms are proven.
Workload-fit risk: No benchmark proves an alternative configuration can run the workload.

PROVIDER AND BUYER VERIFICATION

- Confirm buyer rate, term, and capacity; exact 8-GPU dates, geography, and orderability.
- Collect network, storage, egress, support, tax, deposit, reservation, software, interconnect, and benchmark requirements.

PROPOSED RECHECK AND NEXT REVIEW

- Recheck for a comparable offer, AWS terms or capacity changes, idle variance, or workload changes.
PROVISIONAL - NOT REAFFIRMED. Next review October 27, 2026. No monitoring is active.

EVIDENCE MANIFEST

- August 28 capture, August 24 evidence: Advisor scope/AWS baseline; Audit annualization/12% model; held \$3.90 signal. Some rows excluded.

Producers: Advisor and Audit provide dated evidence; they do not author, approve, or seal.	Renderer: no calculation or offer selection; renders accepted facts and explicit unknowns.
State: illustrative sample, not a delivered Brief or a recommendation.	Privacy: no customer data; customer records remain inside authenticated AIForge Works unless explicitly authorized.

Illustrative fictional draft; governed confirmation is required. Catalog prices do not prove availability, orderability, capacity, quotes, paid rates, performance, or savings. Findings are non-additive; verify with providers. Captured August 28 from the August 24 snapshot; some rows were excluded; \$3.90 mapping remains unresolved.