COMMERCIAL STRATEGY · EXPLAINER

Token Economics — How It Works

Why the same GPU can earn $2.10 an hour or $9.79 an hour, and what decides which one you get.

- Two meters on identical silicon: reserved GPU-hours, or tokens served.
- One worked 8-GPU node carries every number in this explainer.
- The whole question is who carries idle risk — and what you are paid for taking it.

THE FOUNDATION

Two meters, two risk profiles

A GPU fleet can be monetized on exactly two meters. The choice sets who absorbs unsold capacity.

Dimension	Reserved capacity	Token endpoint
Billable unit	GPU-hour committed	Million billable tokens
Revenue driver	Hours sold × rate card	Traffic density × price per M
Idle risk	Customer	Provider
Upside	Capped at the rate card	Scales with density and mix
Gross margin	Flat, about 47% at $2.10	Rises 28% → 82% with density
Forecastability	High — contracted	Lower — demand-driven

THE REFERENCE NODE

The cost floor you must clear

Cost per node-hour is fixed the moment the node is powered on. Only revenue is variable.

Parameter	Value	Note
GPUs per node	8 × H100-class	One serving node
Cash opex	$0.49 / GPU-hr	Power, colo, operations
Capex	$40,000 / GPU	5-year straight line
Depreciation	$0.913 / GPU-hr	40,000 ÷ (5 × 8,760 hrs)
Fully-loaded cost	$1.403 / GPU-hr	= $11.23 per node-hour
Peak throughput	16,667 tok/s	70B-class model, aggregate
Billable share	59.6%	Output-weighted, net of cache hits
Blended price	$2.19 / M tokens	Input + output blend

$11.23

Fully-loaded cost per node-hour — cash opex plus depreciation, incurred whether the node is busy or idle.

THE ARITHMETIC

What one node bills at full density

16,667 tok/s × 3,600 s × 59.6% = 35.8M billable tokens per hour · × $2.19/M = $78.32 per node-hour.

- Token meter at 100% density: **$78.32 / node-hr** — $9.79 per GPU-hour.
- Reserved capacity on the same node: 8 × $2.10 = **$16.80 / node-hr**.
- The gap is 4.7× — and that gap is entirely payment for absorbing density risk.
- Full density is a ceiling, not a plan. Real fleets run partly loaded, so the honest question is: how loaded is loaded enough?

4.7×

Token revenue vs reserved revenue on the identical node at full traffic density.

THE TWO THRESHOLDS

Breakeven at 14.3%, crossover at 21.5%

Two densities decide every meter assignment. Everything else is negotiation.

- **Cost breakeven — 14.3% density.** $11.23 ÷ $78.32. Below it the node loses money on a fully-loaded basis.
- **Reserved crossover — 21.5% density.** $16.80 ÷ $78.32. Above it tokens beat the take-or-pay rate card.
- Between 14.3% and 21.5% the node is profitable but you would have been better off selling the hours.
- Above 21.5%, every incremental point of density is margin the rate card could never have earned.

14.3%
Cost breakeven density

21.5%
Crossover vs $2.10 reserved

THE DENSITY LADDER

Margin is a function of density, not of price

Same node, same rate card, same cost. Only traffic density moves — and margin swings 54 points.

Density	Tokens/s	Revenue / node-hr	Gross margin	Achieved $/GPU-hr	vs $2.10
20%	3,333	$15.66	28%	$1.96	0.93×
30%	5,000	$23.50	52%	$2.94	1.40×
40%	6,667	$31.33	64%	$3.92	1.87×
50%	8,334	$39.16	71%	$4.90	2.33×
60%	10,000	$46.99	76%	$5.87	2.80×
70%	11,667	$54.82	80%	$6.85	3.26×
80%	13,334	$62.66	82%	$7.83	3.73×

THE LEVERS

Four levers move token margin

In order of leverage — and the first one is not price.

- **1. Traffic density.** Routing, batching, autoscaling, multi-tenant packing. Every point of density is pure margin against a fixed cost floor.
- **2. Billable share.** Output weighting, cache-hit accounting, honest metering of what you actually charge for.
- **3. Throughput per node.** Quantization, speculative decode, continuous batching — more tokens per second is a cost reduction per token.
- **4. Blended price and mix.** Steering traffic toward higher-value models and premium latency tiers.

THE DECISION RULE

Which fleet slice goes on which meter

You do not choose one meter. You assign each slice of the fleet to the meter that fits its demand shape.

Fleet slice	Demand shape	Meter	Why
Anchor tenants	Committed, predictable	Reserved take-or-pay	Locks the cost floor and finances the build
Shared inference	Bursty, many tenants	Tokens	Multiplexing lifts density well past 21.5%
Private endpoints	Sustained, single tenant	Reserved + token floor	Isolation is the product; floor protects margin
Fine-tuning / batch	Schedulable, latency-tolerant	Reserved off-peak	Backfills troughs the token layer cannot
Residual capacity	Whatever is left	Spot / RL-priced	Any revenue above cash opex beats idle

SYNTHESIS

What to take away

Three sentences you can carry into a pricing review.

- The cost floor is fixed and the crossover is low — 21.5% density beats a $2.10 take-or-pay rate card.
- Above the crossover, margin is bought with load and metering discipline, not with price increases.
- Run both meters deliberately: reserved capacity to finance and de-risk the asset, tokens to harvest the upside it was always capable of.

2.8×

Achieved rate vs reserved at 60% density — the realistic operating target for a well-routed shared fleet.