COMMERCIAL STRATEGY · TOKEN ECONOMICS

Token Economics — A Primer

How selling tokens differs from selling GPU-hours, where the crossover sits, and how to decide which fleet slice belongs on which meter.

August 2026 · Illustrative figures built from public benchmarks

1. The two meters

A GPU fleet can be monetized on exactly two meters, and the choice determines who carries which risk.

Reserved capacity (take-or-pay)

The customer buys time: 8 GPUs × $/GPU-hr, committed. Revenue is deterministic and contracted, the customer carries idle risk, and provider upside is capped at the rate card no matter how productive the silicon turns out to be.

Tokens (consumption)

The provider sells inference output priced per million tokens and multiplexes one physical node across many tenants. Revenue now scales with traffic density rather than hours sold — and the provider, not the customer, carries the density risk. That single transfer of risk is the whole subject of this primer.

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

2. The reference node

Every number in this primer stands on one worked node so the arithmetic can be checked end to end.

Parameter	Value	Note
GPUs per node	8 × H100-class	One serving node
Reserved rate	$2.10 / GPU-hr	Take-or-pay anchor
Cash opex	$0.49 / GPU-hr	Power, colo, ops

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

At 100% traffic density the node bills 16,667 × 3,600 × 0.596 = 35.8M billable tokens per hour, which at $2.19/M is **$78.32 per node-hour**, or $9.79 per GPU-hour. Reserved capacity on the same node bills 8 × $2.10 = $16.80 per node-hour. That gap — 4.7× at full density — is the prize, and the density risk is the price of admission.

3. The two thresholds that matter

Density is the share of peak serving throughput the node actually delivers over the billing period. Two thresholds govern the decision.

Threshold	Level	Meaning
Cost breakeven	14.3% density	Token revenue covers fully-loaded cost ($11.23/node-hr). Below this the node loses money on a token meter.
Reserved crossover	21.5% density	Token revenue equals the $16.80/node-hr take-or-pay line. Above this, tokens beat capacity on the same iron.

The crossover is deliberately low, which is the counter-intuitive finding: a token endpoint only needs about a fifth of its serving capability to be busy to out-earn a fully committed reserved contract. But the crossover is a break-even on *expected* revenue, not a decision rule. Apply a haircut for density and price risk before committing a fleet slice: a 30% haircut moves the crossover to roughly 31% density, and a 50% haircut to roughly 43%. Sell tokens where you can defend the haircut-adjusted level, not the raw one.

4. The density ladder

The same node, read across achievable densities. Achieved $/GPU-hr is the number to compare against the $2.10 rate card.

Density	Tok/s	Token rev /node-hr	Achieved $/GPU-hr	Gross margin	vs res.	Read
20%	3,333	$15.66	$1.96	28.3%	0.93×	Below crossover — keep on capacity
30%	5,000	$23.49	$2.94	52.2%	1.40×	Marginal; needs a demand floor

Density	Tok/s	Token rev /node-hr	Achieved $/GPU-hr	Gross margin	vs res.	Read
40%	6,667	$31.33	$3.92	64.2%	1.86×	Token meter clearly wins
50%	8,333	$39.16	$4.89	71.3%	2.33×	Target band for shared endpoints
60%	10,000	$46.99	$5.87	76.1%	2.80×	Steady state for a healthy endpoint
70%	11,667	$54.82	$6.85	79.5%	3.26×	Add capacity before quality slips
80%	13,333	$62.65	$7.83	82.1%	3.73×	Latency SLO becomes the constraint

Margin rises steeply and then flattens because cost per node-hour is fixed while revenue is linear in density. Almost all of the economic gain is captured between 20% and 60%; past 70% the binding constraint stops being margin and becomes the latency SLO — queueing delay, not profitability, sets the ceiling.

5. What actually moves token margin

Four levers, in order of leverage per unit of effort.

- **Density (traffic per node).** Linear in revenue with zero cost change. Consolidate tenants onto fewer models and fewer nodes before buying more silicon; two nodes at 60% beat three at 40%.
- **Billable share.** The 59.6% assumption is a lever, not a constant. Prompt-cache hits and speculative decoding raise throughput but are not always billable — define what is billable in the price book before engineering optimizes against it.
- **Throughput per node.** Batching, quantization, paged attention and a better serving stack raise peak tok/s on unchanged hardware, which lowers both thresholds. Every 10% throughput gain moves the crossover down roughly two points of density.
- **Blended price and mix.** Output tokens cost more to produce than input tokens; long-context and reasoning workloads shift the blend. Price by model class and context window, not one flat $/M.

6. Deciding which slice goes on which meter

Fleet slice	Meter	Why
Anchor enterprise commitments	Reserved, take-or-pay	Contracted revenue underwrites the debt and fixes the capacity floor
Shared inference endpoints	Tokens	Multiplexed density is 2–3× the reserved rate above crossover
Private / dedicated endpoints	Reserved + token overage	Isolation is the product; tokens capture upside above the floor

Fleet slice	Meter	Why
Fine-tuning and batch jobs	Reserved or spot	Bursty and latency-tolerant — fills troughs the token layer cannot
Residual / uncommitted capacity	Tokens, dynamically priced	Perishable; any density above 14.3% beats an idle node

7. Risks and controls

- **Density risk.** Revenue is demand-driven. Hold a reserved floor across the fleet so a bad quarter of traffic cannot break debt service.
- **Metering integrity.** A token business is only as good as its token counter. Meter both legs at the serving layer, reconcile three ways (usage → rated → invoiced), and make unmatched usage a blocking control.
- **Price compression.** Public $/M-token prices fall faster than hardware depreciates. Re-underwrite the price book quarterly and hedge the capacity leg where a futures market exists.
- **SLO cost.** Chasing density past 70% degrades latency and triggers credits, which shows up as revenue leakage rather than as a cost line. Cap density at the SLO, not at the margin optimum.

8. The one-line conclusion

On identical hardware, a token meter beats a $2.10 take-or-pay contract above roughly 21.5% traffic density and reaches 2.8× the achieved rate at 60% — so the commercial question is never “tokens or capacity” but **how much of the fleet can be defended above the haircut-adjusted crossover**, with the rest anchored on take-or-pay.

All figures are illustrative — synthetic values built from public benchmarks, internally consistent but not any specific company's economics. The live version of this model, with adjustable density, price and cost assumptions, is the Token economics tab of the cockpit.