COMMERCIAL STRATEGY · PRICING SCIENCE

Finding the Optimal Combination of Pricing Levers

Maximizing gross margin and revenue-weighted utilization across a depreciating GPU fleet.

August 2026 · Illustrative figures built from public benchmarks

Gross margin on a GPU cloud is not set by a rate card. It is the output of four interacting commercial levers applied to different slices of the same physical fleet: how much capacity is sold as certainty, how much is multiplexed into tokens, how the perishable residual is priced and hedged, and how tightly the discount waterfall is governed. Optimizing any one lever in isolation moves revenue between slices rather than raising fleet margin.

This paper sets the unit cost floors from the cost stack, then walks the four levers in the order they must be decided, with the arithmetic behind each threshold: the 60–75% take-or-pay anchor, the 21.3% token crossover density, the $2.10 hedge strike against a $1.70 spot clear, and the 237,442-token outcome-versus-metered crossover.

Section	Decision it supports
1. Cost floors	What every lever must clear before it can be called margin
2. Take-or-pay anchor	How much capacity is sold as certainty
3. Token multiplexing	Which slice should serve tokens instead of hours
4. RL residual & hedging	How the perishable tail is priced and de-risked
5. Packaging & waterfalls	Outcome vs metered, and what the desk may concede
6. Target lever mix	The blended fleet target: 82–88% RWU at 50–58% margin

Companion to the P&L primer (/primer), the strategy & GTM document (/strategy), the compute futures memo (/futures) and the billing cockpit (/). The live version of this paper is at /levers.

1. The four levers & cost floor baseline

No lever can be evaluated without the floors it must clear. From the cost stack in the primer, per GPU-hour:

Cost basis	$ / GPU-hr	Composition
Cash opex floor	$0.485	Power/cooling $0.243 + network $0.082 + software $0.094 + support $0.066
Fully loaded, ex-interest	$1.495	Cash opex $0.485 + 5-yr straight-line depreciation $1.010
Fully loaded, inc. interest	$1.899	$1.495 + $0.404 debt service @ 70% LTV, 8% interest

Gross margin % = (achieved $/GPU-hr − cost-to-serve $/GPU-hr) / achieved $/GPU-hr.

- **Lever 1 — contract mix.** The share of installed capacity sold as multi-year take-or-pay certainty versus held merchant.
- **Lever 2 — packaging.** Whether a given slice is sold as GPU-hours or multiplexed into tokens, endpoints and outcomes.
- **Lever 3 — dynamic price.** The sequential price path on residual capacity, plus the financial hedge on hours that would otherwise perish.
- **Lever 4 — discount governance.** The list-to-realized waterfall: what the desk is allowed to concede, and under what control.

The order of decision

The levers are decided in sequence, because each one sizes the slice the next one operates on: set the take-or-pay floor (60–75% at $2.05/hr), overlay token multiplexing where density exceeds 21.3%, price and hedge the residual, then govern the discount waterfall so leakage stays under 0.5%.

2. Step 1 — anchor 60%–75% of capacity in take-or-pay

Target rate: list $2.35, realized $2.05, hard floor $1.95 per GPU-hour. The role of this lever is not margin maximization — it is transferring utilization risk to anchor customers so that fixed depreciation is covered before any merchant hour is sold.

1,024-GPU cluster at 100% contract coverage ($45.3M capex)

Line	Value	Note
Gross revenue	$18.84M / yr	$2.05/GPU-hr realized
Cash EBITDA	$14.49M	76.9% margin over the $0.485 cash floor
Depreciation	−$9.06M	5-yr straight line, $1.010/GPU-hr
Interest	−$2.54M	70% LTV @ 8%
Pre-tax income	$2.89M	15.3% pre-tax net margin

Rule: never let take-or-pay drop below 60% of installed capacity. Below that, fixed depreciation of $1.01/GPU-hr is spread across too few contracted hours and dilutes fleet gross margin regardless of how well the merchant book prices.

3. Step 2 — overlay token serving on the flex capacity

Token serving is a multiplexing arbitrage: the same node is sold many times over to concurrent requests, so revenue per GPU-hour scales with traffic density rather than with the rate card.

Input	Value
Output tokens	$2.19 / M tokens
Input tokens	$0.55 / M tokens
Cached input	$0.055 / M tokens
Internal capacity transfer cost	$2.10/GPU-hr = $16.80/node-hr (8×H100)
Throughput, 70B class	~10,000 output tokens/sec = 36M tokens/node-hr
Revenue at 100% density	$2.19/M × 36M = $78.84/node-hr ($9.85/GPU-hr)

Crossover density

crossover = $16.80 base cost / $78.84 peak revenue ≈ 21.3%

At 60% traffic density the node earns $47.30/hr ($5.91/GPU-hr) — a 75.6% gross margin against roughly 33% on the same silicon sold as reserved capacity.

Rule: route uncommitted capacity into token serving only when sustained traffic density is expected to exceed 21.3%. Below that threshold the endpoint destroys value relative to simply renting the node.

4. Step 3 — price the residual dynamically and hedge it

The remaining 15–25% of the fleet is spot and on-demand. Pricing here is sequential, not static: a customer priced out today defers and may return, so today's discount destroys tomorrow's price. That is the structure a contextual bandit exploits in the cockpit's RL simulation tab.

Cohort	Willingness to pay	Elasticity	Patience
Anchor AI lab (burst)	$2.90	0.9	0.25
Enterprise / BFSI	$3.10	0.6	0.55
Developers / self-serve	$1.90	2.4	0.70
Batch / research	$1.60	1.6	0.85

- **Dynamic policy.** Hold price at $2.40–$2.60 while the deferred-demand backlog is high; discount to $1.60–$1.80 only into soft demand weeks, to clear perishable hours above the $0.485 cash opex floor.
- **Hedge the tail.** Short compute index contracts on 30–35% of uncommitted volume at a $2.10 strike. If spot clears at $1.70 the hedge returns +$0.40/hr, holding cash EBITDA margin above 70%.
- **Never hedge the core.** Take-or-pay hours are already contractually fixed; hedging them converts a stable book into a speculative one.

5. Step 4 — govern packaging and discount waterfalls

Outcome versus metered pricing

Pricing an agent workflow at $0.52 per resolved task, against a token cost of $0.63/M and a metered list of $2.19/M:

crossover = ($0.52 / $2.19) × 1,000,000 = 237,442 tokens per task

Sell outcome-based pricing for agent tasks consuming under 237k tokens, where it carries the higher margin; revert to metered token rates for long-context tasks above that line.

Waterfall discipline

Stage	Value	Control
List	$141.6M	Governed price book, versioned
Contracted	$113.9M	Approved standard discount bands
Realized	$111.8M	Concessions, credits and leakage

Cap non-standard concessions below 3% and enforce four-eyes database gates (control C-11) to prevent unapproved discounting. Target billing leakage below 0.5%.

6. Target optimal lever mix

Fleet slice	Commercial wrapper	Target realized rate	Utilization target	Gross margin
60%–75%	Reserved take-or-pay	$2.05 / GPU-hr	100% (take-or-pay)	27.1% (ex-interest)
15%–25%	Token-as-a-service	$2.19 / M tok ($5.91/GPU-hr eq.)	> 50% density	68%–75%
10%–15%	On-demand / RL spot / hedged	$1.70–$2.60 + $2.10 strike	65%–85%	45%–60%
Blended fleet	Multi-tier portfolio	$2.40–$2.70 / GPU-hr eq.	82%–88% RWU	50%–58%

Where to inspect and model this in the cockpit

- **Dynamic policies & cohort elasticity.** The RL simulation tab on the cockpit (/).
- **Cluster hedge & payback sensitivity.** The compute futures page (/futures).
- **List-to-realized discount waterfalls.** The pricing & packaging tab on the cockpit (/).
- **Full unit-cost economics and worked P&Ls.** The P&L primer (/primer), sections 4 and 5.

All figures are illustrative — synthetic data built from public benchmarks, consistent with each other but not any specific company's economics.