GPU.bot

WHITE PAPER · V2.0 · AUGUST 2026

The Compute Spread.

An AI-operated brokerage for the GPU economy

Identical NVIDIA H100s rent for $1.49 to $6.98 per hour across public clouds — a 23× spread on a commodity input to the fastest-growing market in technology. GPU.bot is a demand-side brokerage that captures that spread with a workforce of AI agents running on the Claude API: scanning 40+ providers, matching workloads, generating deploy configs, and charging 20% of documented savings. Software margins. No inventory.

23×	42	$26.1B	<$1
H100 price spread	providers tracked	GPUaaS by 2031	cost to broker a deal

01 · EXECUTIVE SUMMARY

Sell the spread, not the silicon

The market for rented GPU compute is large, growing fast, and pricing itself badly. In 2026, published rates for an identical NVIDIA H100 range from $1.49 to $6.98 per GPU-hour across public providers — and as low as $0.38 on spot markets — a spread of up to 23× on a commodity input. Specialized GPU clouds routinely undercut hyperscalers by 50–75%, yet most buyers still default to the expensive incumbents because comparing 40+ providers, hundreds of SKUs, and volatile spot markets is a full-time job nobody has.

GPU.bot makes it a full-time job that software has. It is a demand-side compute brokerage whose entire operating core — price collection, normalization, workload matching, deploy-config generation, price alerts, market reporting, and first-line support — runs as scheduled agent loops on the Claude API. The company owns no GPUs, holds no inventory, and takes no infrastructure risk. It sells three things: a free live price index that acquires users, paid tooling that converts them, and brokered savings that monetize serious spend at 20% of documented savings.

The result is a business with software gross margins on top of an infrastructure boom: the GPU-as-a-service market is projected to grow from roughly $7.4B in 2026 to $26.1B by 2031 (28.7% CAGR), inside an AI infrastructure market exceeding $100B and hyperscaler capex of $725B in 2026 alone. A brokerage that intermediates even a fraction of a percent of that flow, at a marginal cost per transaction measured in cents, is among the highest-leverage businesses an AI-native company can run today — because the AI is not a feature of the product; it is the workforce.

> **“The AI is not a feature of the product. It is the workforce.”**

And the brokerage is only the first rung. The endgame: publish the number everyone cites — the GBX-H100 reference rate — standardize the unit everyone quotes, and become the venue where compute clears. Rate, unit, venue: the CME playbook, applied to the defining commodity of the decade.

02 · THE MARKET

A $26B market priced like a bazaar

AI compute is the defining industrial buildout of the decade. Amazon, Google, Meta, and Microsoft confirmed combined 2026 capital expenditure of roughly $725B, up from $410B in 2025, with GPUs the dominant hardware line item. Downstream of that buildout sits the rental market GPU.bot operates in: GPU-as-a-service, estimated at $7.4B in 2026 and projected to reach $26.1B by 2031 at a 28.7% CAGR, within an AI infrastructure market of $101B in 2026.

Supply has fragmented faster than buyers can track it. Independent trackers now monitor 54 distinct GPU cloud providers: hyperscalers (AWS, Google Cloud, Azure, Oracle), specialty AI clouds (CoreWeave, Lambda, Crusoe, Nebius), marketplaces (RunPod, Vast.ai, Shadeform), bare-metal hosts, and regional clouds. One comparison site counts 247 live H100 listings across 42 providers. Each publishes its own SKUs, regions, interconnects, billing increments, and reliability tiers. There is no consolidated tape, no NBBO, no ticker. It is a commodity market trading like a bazaar.

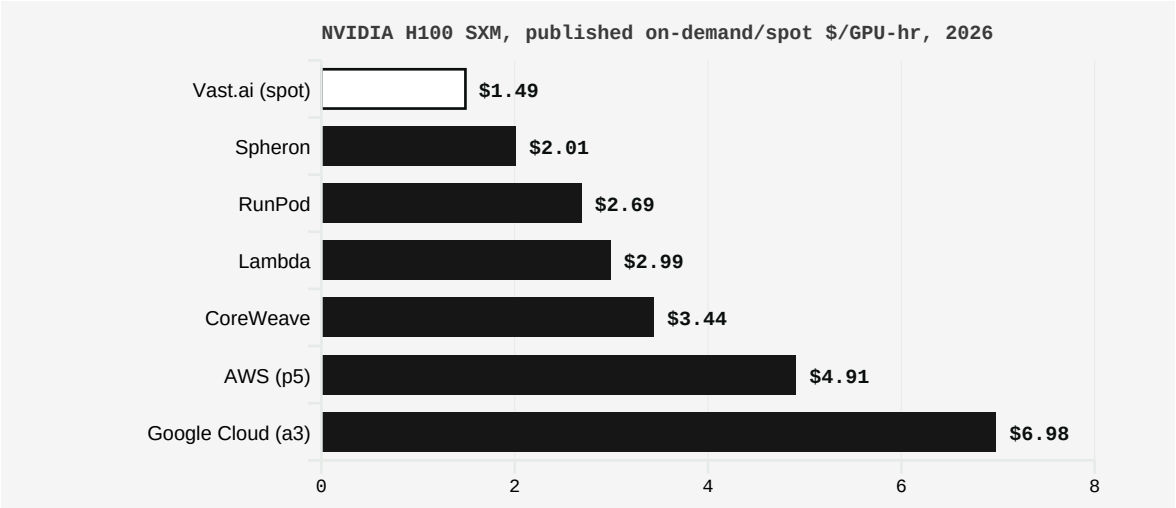

Figure 1 — Published H100 pricing across representative providers, 2026. Sources: getdeploying.com, gpuperhour.com, provider rate cards. Spot listings reach $0.38–$1.49/hr.

03 · THE INEFFICIENCY

23× dispersion, and it isn't closing

Price dispersion this extreme persists for structural reasons. Switching feels risky, so buyers anchor to hyperscalers whose brands they trust, paying a 2–4× premium for identical silicon. Information is fragmented across dozens of frequently-changing pricing pages. And volatility adds a time dimension: AWS H100 spot pricing has swung from $1.52 to $0.74 per GPU within two weeks, while other providers held flat, and analysts report 40–60% cost reductions available to batch-tolerant workloads that simply route intelligently across regions and hours of the day.

Every prior commodity market with this profile — equities before consolidated tape, ad inventory before programmatic exchanges, freight before digital brokers — converged on the same answer: an intermediary that standardizes price discovery and takes a spread for execution. GPU compute is at the pre-intermediary stage. The dispersion is the business plan.

MARKET SIGNAL	VALUE (2026)	IMPLICATION
H100 price range, on-demand/spot	$1.49 – $6.98/hr	Up to 23× spread incl. spot
Market average H100 rate	$3.57/hr	Most buyers overpay vs. best qualified offer

Specialty cloud vs. hyperscaler	50–75% cheaper	Brand-anchoring premium is capturable
AWS H100 spot swing (2 weeks)	$1.52 → $0.74	Timing alone is worth ~50%
Providers / live H100 listings	42 / 247	Too many venues to compare manually
GPUaaS market, 2026 → 2031	$7.4B → $26.1B	Spread volume compounds 28.7%/yr

Table 1 — The dispersion opportunity in numbers. Sources listed in §9.

04 · THE PRODUCT

An index, a matcher, and a limit order for compute

GPU.bot presents as three products sharing one data spine. The **Index** is a free, continuously updated price tape: every tracked provider's GPU offerings normalized to $/GPU-hour with VRAM, interconnect, region, and reliability attached. It is the top of the funnel and the market's missing public good. The **Matcher** takes a workload described in plain English — “fine-tune an 8B model, ~200 GPU-hours, interruptions acceptable” — compiles it into hard requirements, ranks every live listing, and returns the three cheapest qualified offers with a ready-to-run SkyPilot or Terraform config for the winner. The **Watchdog** is a limit order for compute: “alert me when 8×H100 clusters print under $2.00/hr in us-east” — held open by agents that monitor the tape around the clock.

Deliberately absent: GPU.bot never touches customer workloads, weights, data, or cloud credentials. It generates configs; customers execute them in their own accounts. This keeps the security surface near zero and the trust story simple — GPU.bot is a broker and an information business, not an infrastructure operator.

05 · ARCHITECTURE

Four agent loops on the Claude API

The design constraint from day one: every recurring operation must be executable by a commercially available AI agent today — no speculative capabilities, no human assembly line hiding behind the bot. The system decomposes into four scheduled loops.

LOOP	CADENCE	WHAT THE AGENT DOES
Scout	Every 15 min	Fetches provider pricing pages/APIs; parses messy HTML, JSON, and PDFs into one normalized s
Match	On request	Compiles plain-English job specs into constraints; ranks live listings; drafts quote with savings vs. m
Deploy	On acceptance	Generates and validates SkyPilot/Terraform config for the winning provider; returns one-command l
Watch	Continuous	Holds price limit orders; fires alerts and re-quotes when targets print; monitors brokered deals for fa

Table 2 — The operating core. Two auxiliary loops — a report writer that turns index data into the weekly GPU Market Report, and a support agent that drafts responses to inbound mail — complete the headcount-free operating stack.

Each loop is a narrow, verifiable task with structured inputs and outputs — precisely the regime where current-generation models are reliable. Parsing a pricing page into a schema, compiling a requirement, filling a config template: these are bounded problems with checkable answers. Guardrails are structural rather than aspirational: agents write to a versioned data store, price outliers beyond statistical bounds are quarantined pending a second-pass check, generated configs are linted and dry-run before delivery, and anything touching money or contracts routes to a human. Humans in the company do exactly three jobs: sign agreements, hold the bank account, and audit samples of agent output. Everything else is inference.

06 · BUSINESS MODEL & UNIT ECONOMICS

Cents of inference against thousands in savings

Revenue stacks in four tiers. **Index (free)** acquires users and makes GPU.bot the citation of record for GPU pricing. **Pro ($49/month)** sells unlimited matching, limit orders, config generation, and a 10k-call API to individual builders and small ML teams. **Broker (20% of documented savings)** serves accounts spending $10k+/month, where agents negotiate, arbitrage spot windows, and manage failover — with fees charged only against an auditable savings baseline. **Data ($499/month)** sells the historical tape, volatility surfaces, and capacity signals to funds and analysts pricing the AI buildout — the Bloomberg seat for the GPU economy.

The cost side is where the model breaks from every brokerage before it. A full market scan — 42 providers parsed to schema — costs a few dollars in inference; the entire always-on data plane runs under roughly $2,000/month. A match request costs a few cents. Executing a brokered deal costs well under $1 of API usage against hundreds of dollars of fee.

ILLUSTRATIVE BROKERED DEAL

Workload	8×H100, 72 hours	576 GPU-hours
At market average ($3.57/hr)	$2,056	what the buyer would have paid
At best qualified offer ($1.87/hr)	$1,077	what the buyer pays instead
Documented savings	$979 (48%)	audit trail attached to invoice
GPU.bot fee (20% of savings)	$196	buyer keeps $783
Inference cost to execute	< $1	gross margin on the fee: >99%

Table 3 — Unit economics of a single mid-sized brokered deal, using 2026 published rates.

Subscription economics behave the same way: a heavy Pro user consumes perhaps $3/month of inference against a $49 price — a ~94% gross margin that improves as models get cheaper, a cost

curve that has fallen consistently every year since 2023. GPU.bot's input costs deflate on the same curve that inflates its market: cheaper inference means cheaper operations for GPU.bot and more AI workloads needing GPUs for everyone else. The business is long AI on both sides of its income statement.

07 · GO-TO-MARKET

A flywheel that publishes

The index is the growth engine. Public, free, continuously fresh pricing data is inherently linkable — every “what does an H100 cost” search, newsletter citation, and pricing debate is an acquisition channel that compounds without spend. On top of it, the report-writer agent publishes the weekly GPU Market Report — spread movements, capacity signals, notable arbitrage windows — generated from GPU.bot's own tape. Content marketing with zero marginal authorship cost, in a niche where every reader is a qualified buyer.

Readers set a free price alert; the first alert that fires demonstrates concrete money left on the table; Pro converts the demonstration into tooling; accounts whose spend crosses $10k/month graduate to brokered deals, where the audited savings baseline makes the 20% fee an easy yes. Providers, meanwhile, gain qualified demand when they price competitively on the index — which recruits supply-side cooperation and partner API access, which improves the tape, which sharpens the index. Both sides of the market strengthen the asset the company monetizes.

The strategic priority inside this flywheel is citation. Every dashboard, newsletter, investor deck, and cloud contract that quotes “GBX-H100” hardens the index into a benchmark — and benchmarks are winner-take-all: once a market quotes a reference rate, no single participant can unilaterally switch away from it. The index is given away precisely because it is the moat.

08 · RISKS, HONESTLY

What could kill this, and what we do about it

The spread compresses. If transparency (partly ours) narrows dispersion from 23× toward 2–3×, savings-share revenue shrinks. Mitigation: even mature commodity markets sustain intermediaries on volatility, timing, and complexity — and the Data tier monetizes transparency itself. The more efficient the market becomes, the more valuable its tape.

Data access tightens. Providers could obscure pricing or block collection. Mitigation: public pricing is how challengers win demand from incumbents — the competitive logic that created the dispersion also sustains disclosure. Partner APIs and provider-submitted feeds (free listing, verified badge) convert adversarial scraping into cooperative supply.

Agents make mistakes. A bad parse could misprice a quote. Mitigation: structural verification — statistical quarantine of outliers, second-pass checks, config dry-runs, and human sign-off on every

contract-touching action. Error rates on bounded parsing tasks are already low and fall with each model generation; the audit trail is published with every deal.

Platform concentration. The workforce runs on one vendor's API. Mitigation: every loop is prompt-plus-schema with no exotic dependencies; the architecture is model-portable by construction, and multi-model redundancy for the Scout loop is on the roadmap.

Incumbent competition. Shadeform aggregates supply; SkyPilot orchestrates workloads; trackers publish comparisons. None combine a public tape, plain-English brokerage, and savings-share economics in one AI-operated stack — and the operating-cost asymmetry means a headcount-based copy of GPU.bot runs at a structural disadvantage to the original.

09 · THE LADDER

Rate. Unit. Venue.

GPU.bot's trajectory is the one every major commodity market has already walked — compressed, and run by agents. Each rung is a working business on its own; each funds and legitimizes the next; none depends on the next succeeding.

RUNG	STATUS	WHAT IT IS	REVENUE
01 · Index	Live	GBX-H100 reference rate — free, published, cited	Acquisition
02 · Tape	Live	Tick history + volatility surfaces; cannot be backfilled	Data, $499/mo
03 · Contract	Draft	One standard unit: GBX-H100.99.USE	—
04 · Exchange	2027	Resting orders both sides, cleared matches	Basis points
05 · Derivatives	2028	Futures and forwards settled on the daily print	Clearing + data

Table 4 — Each rung compounds the moat of the one below it.

The draft standard unit — **GBX-H100.99.USE**: one GPU-hour of H100-class silicon at a ≥99% uptime tier in us-east, marked every 15 minutes, settled daily at 00:00 UTC against the index — is what turns comparison into trading. Once buyers and sellers quote in one unit, resting orders can meet at a venue, matches can clear for basis points, and futures can settle against the daily print. Exchanges routinely operate at 70%+ gross margins behind network effects that deepen with every participant, and AI companies have an acute, unhedged exposure to compute costs that only index-settled contracts can close. The entity that owns the settlement price owns the market. That is the position the tape is quietly accumulating.

> **“Whoever owns the settlement price owns the market.”**

SOURCES & DATA

H100 pricing, spread, listings count: getdeploying.com/gpus/nvidia-h100 · gpuperhour.com · intuitionlabs.ai/articles/h100-rental-prices-cloud-comparison · spheron.network/blog/gpu-cloud-pricing-comparison-2026

B200/GB200 pricing: getdeploying.com/gpus/nvidia-b200 · thundercompute.com/blog/nvidia-b200-pricing

Market sizing: Mordor Intelligence, GPU-as-a-Service Market ($7.38B 2026 → $26.09B 2031, 28.73% CAGR) · Mordor Intelligence, AI Infrastructure Market ($101.17B 2026) · Fortune Business Insights, GPUaaS

Hyperscaler capex ($725B combined 2026): company guidance as compiled by Mordor Intelligence / IDC

Spot volatility & multi-cloud savings (40–60%): cast.ai GPU Price Report 2026 · spheron.network GPU spot arbitrage analysis

Provider landscape (54 tracked): gputracker.dev · SkyPilot GPU Compass (blog.skypilot.co/gpu-compass)

All prices are published rates as of mid-2026 and fluctuate; figures are illustrative of market structure, not quotes.

GPU.bot · hello@gpu.bot · The index is free. The savings are yours. The spread is ours.