Bank of America and Citi Invest in Silicon Data Read the announcement →

The GPU Forward Curve: Pricing the Curve, Not Just the Spot

Why a GPU forward curve is the missing primitive that turns compute rental into a tradeable, hedgeable asset class for financial institutions.



Written by Carmen Li
Founder at Silicon Data

Industry May 1, 2026 13 Mins Read

Summarize with AI

An asset class that scaled before its pricing infrastructure did

Crude oil had a forward curve before NYMEX cleared its first WTI future. Short-term interest rates had a term structure long before SOFR swaps reached institutional volumes. Electricity, natural gas, shipping freight, even iron ore — each of these markets produced a continuous, methodology-backed view of where prices were expected to go *before* the derivatives built

on top of them found liquidity. The curve came first. The instruments came second.

AI compute has been flying blind in the future. The asset has financialized at extraordinary speed: over $176 billion in GPU-collateralized infrastructure now sits on hyperscaler and neocloud balance sheets, private credit funds are underwriting neocloud debt, and insurers have begun exploring parametric products tied to compute cost exposure. Rental is a multi-billion-dollar annual market distributed across dozens of providers. And yet, its core foundation remains half-built — until recently, there was no published forward curve for any of it.

This is the missing piece. The question worth asking is not when GPU futures will develop. It is why the pricing primitive that every other commodity market treats as a precondition has taken this long to arrive — and what becomes possible now that it has.

What it has cost the market to operate without one

The absence of a forward curve is not a technical inconvenience. It is an underwriting problem, a hedging problem, and a product-design problem that has been quietly distorting how capital is deployed into AI infrastructure.

Consider the lender underwriting a $400M credit facility to a neocloud. The revenue line in the model depends on expected rental rates for H100 and B200 capacity 12, 24, and 36 months out. Without a market-calibrated forward curve, where does that number come from? It comes from the borrower's own projection. The party least incentivized to stress-test the downside case is the one supplying the forecast. Every major infrastructure lender privately knows this. Most have no better option.

The hedging problem is structurally similar. A compute-heavy AI company that wants to cap its rental cost exposure at a 12-month forward rate needs a counterparty willing to quote that rate. The counterparty needs a mark. Without a published forward curve, there is no mark, which means the quote is either punitive or non-existent. Trading desks that have fielded these requests — and several have — have consistently turned them away for the same reason. The hedge that would be written on any mature commodity cannot be written on compute.

The product-design problem compounds the first two. Parametric insurance — the kind of contract that pays out when an observable index deviates from expectation — requires a published "expectation" to reference. A policy triggered by GPU rental rates falling 20% below the curve needs a curve to compare against. Without one, the product is a conversation. With one, it is a contract.

In each case, the consequence is the same. Risk that the market *could* be pricing collectively is instead being priced by whichever party has the strongest narrative, the deepest balance sheet, or the least exposure to being wrong. That is not a functioning market. That is a market waiting for its reference rate.

The thesis

A GPU forward curve is not a product feature. It is the missing primitive that converts GPU rental from an operational cost line into a tradeable, hedgeable asset class. Without it, compute remains infrastructure spend. With it, compute becomes a market.

Why this methodology, from this source

The construction is not novel in concept. A forward curve derived via no-arbitrage is the same mathematical backbone used in every sovereign yield curve, every commodity futures curve, and every FX forward book since the 1970s. The inputs change. The derivation does not. Applying the method to GPU rental rates is therefore not an innovation in financial engineering so much as an act of recognizing that the asset class is now mature enough to deserve the treatment.

What matters is whether the party publishing it has the market access and the domain standing to be trusted as a reference. Silicon Data's spot index suite covering A100, H100, and B200 is already distributed through Bloomberg and Refinitiv and widely cited in industry and press. The company is backed by DRW and Jump Trading — two firms whose core competence is exactly the kind of pricing infrastructure that makes tradeable markets possible. Partnerships with NVIDIA and TSMC provide upstream visibility into the supply side of the curve. And the rental data itself spans 95% of neocloud providers and 100% of major hyperscalers, which closes the gap between "a price" and "the market clearing price."

None of that makes the curve correct by fiat. It makes it auditable. That is the relevant standard for a settlement reference.

How to read the GPU forward curve

The curve delivers two related signals. Both need to be understood, because they carry different economic meanings.

The **term structure rate** is the all-in hourly rate a buyer would lock in today to rent a given GPU for a fixed forward-starting contract of length T. A 12-month H100 term rate is the rate at which a 12-month contract clears the market right now. It is observable. It is the primary input to the curve.

The **forward rate** is derived from the term structure through the no-arbitrage principle. It answers a different question: as of today, what rate does the market imply for renting that GPU at a point *T* months in the future, for an infinitesimally short duration starting at that future instant? It is not a forecast in the editorial sense. It is the mathematical consequence of the current term structure — the rate at which no arbitrage exists between locking in today and waiting to lock in later. By construction, the forward rate at term zero equals today's spot index.

The more useful thing to internalize is the *shape* of the curve, because the shape carries economic meaning that no single rate can express:

- **Upward sloping (contango)** — the market expects rates to rise. Consistent with anticipated demand growth or supply constraints at longer horizons.
- **Downward sloping (backwardation)** — the market expects rates to fall. Consistent with anticipated supply expansion (new chip generations coming online, hyperscaler buildouts completing) or demand softening.
- **Humped** — near-term rates elevated relative to both short and long tenors. Typical of current supply tightness that the market expects to resolve at a specific horizon.
- **Inverted hump** — a localized dip at an intermediate tenor. A signal of expected supply relief or demand softness concentrated at a particular future moment.

Once a reader can read the shape, the reader can form a view on any future curve without being instructed what to do with it. That is the generative point: the shape is the signal. The absolute level is merely the starting point.

What the curve unlocks that a spot index cannot

A spot index tells a buyer what clearing rates look like today. That is more informative than provider rate cards or broker estimates — a single month's index data can already reveal supply constraints, pricing fragmentation, and regime shifts that a rate card cannot. But it is insufficient for the three applications that define a financialized asset class.

Swap pricing. A 12-month H100 rental swap cannot be marked without a 12-month forward rate. The counterparty quoting the swap needs to know the market's forward expectation; otherwise, the quote is either a speculative bet on future rental prices or a refusal to transact. Historically, most desks have chosen the refusal. With a published forward curve, the swap can be priced against a reference the way any interest-rate or commodity swap is priced — which means the product becomes transactable at institutional size rather than bespoke and friction-laden.

Covenant modeling and credit underwriting. A neocloud's debt service coverage ratio 24 months forward depends on rental revenue 24 months forward. The forward curve replaces the borrower's own projection with a market-calibrated trajectory, and the credit team can stress-test covenants at ±1σ deviations from the curve rather than at arbitrary haircuts. This is not a marginal improvement. It is the difference between underwriting on a narrative and underwriting on a reference.

Parametric insurance and structured products. A policy that pays out when rental rates fall 20% below expectation requires a published "expectation." A structured note that delivers enhanced yield when compute rates stay within a forward-curve-defined corridor requires the corridor to be defined by something other than the issuer's own view. The existence of a daily-published, methodology-backed forward curve is what turns each of these from whiteboard concepts into originatable products.

In each case, the transition is from conversation to contract. The curve is what makes that transition possible.

The curve is observer-independent

This point matters because it is often misunderstood. The forward rate is not Silicon Data's forecast. It is a mathematical derivation from observed term structure rates using standard no-arbitrage methodology. Anyone who disagrees with a forward rate is disagreeing with the collective pricing behavior of the rental market itself, not with any party's opinion about the future.

The term structure inputs are observed across 95% of neocloud providers and 100% of major hyperscalers. Outliers are scrubbed. Specs, interconnect, and region are normalized. The forward derivation is the same calculation a rates desk applies to a sovereign curve or a commodity desk applies to Brent. Given the same inputs, any competent counterparty replicates the same output.

That is the property that makes a rate tradable. A forecast cannot be a settlement reference because forecasts are opinions. A no-arbitrage derivation can be a settlement reference because the construction is reproducible and disclosed. Silicon Data's role is to aggregate the inputs and publish the output daily.

The steel-man objection

The strongest argument against a GPU forward curve is not that the idea is wrong. It is that the underlying market is too heterogeneous to support one. GPU rentals vary by machine spec, interconnect topology, region, contract tenor, cancellation terms, and a dozen other embedded variables. Aggregating across all of that, the objection runs, produces a number that is either too noisy or too abstracted to mean anything.

The objection deserves to be taken seriously, because it is correct about the raw data. A rental contract for H100 SXM5 in a Tier III US data center with 3200Gbps InfiniBand is not the same instrument as a rental for PCIe H100s with Ethernet interconnect in a secondary region. Treating them as fungible would produce a curve that no serious counterparty would reference.

The response is that this is exactly the objection raised against every commodity curve at its inception — an argument Yuhua Yu makes in detail in Three misconceptions in building the financial infrastructure for compute. Brent had — and still has — quality adjustments and basis differentials across delivery points. SOFR had segmentation issues between repo types. Electricity forward curves still contend with zonal basis and transmission constraints. None of those markets solved the heterogeneity problem by waiting for perfect standardization. They solved it through methodology transparency: disclosed normalization, published adjustment factors, documented outlier rules, and continuous refinement as the market matures.

That is the standard the GPU forward curve is built to meet. The question is not whether the methodology is perfect today. It is whether the methodology is disclosed, reproducible, and improving. The alternative — waiting for GPU rental to standardize before publishing a curve — is simply not viable given the pace at which AI compute is financializing on top of an incomplete foundation. Commodities are resolved through methodology, not through homogeneity.

How financial institutions use this now

The practical shift is less a set of steps than a set of frames. Four, specifically, that apply to distinct desks.

Trading desks can use the forward curve as the reference rate for indicative swap quotes on compute exposure. Contango and backwardation regimes

become tradeable basis views — a 12-month contango regime may compress into backwardation as Blackwell supply arrives, and the relative steepness of the curve becomes a positionable thesis. Forward-curve shape is a first-order input to any rental-linked structured product, the same way a yield curve is a first-order input to a rate-linked note.

Credit teams underwriting neocloud or GPU-collateralized debt can substitute the forward curve for the borrower's own revenue projection. Stress-testing at ±1σ or ±2σ deviations from the curve produces a defensible downside case that survives committee scrutiny in a way that "the borrower says rates hold" does not. For revolving facilities with rate-sensitive covenants, the curve — paired with granular spot data on current rental clearing — becomes the natural benchmark against which covenant triggers are modeled.

Insurance originators and parametric product designers gain the settlement reference they have been missing. A policy triggered by a 15% decline in 6-month forward rental rates has a defined, publicly observable trigger. Pricing the premium on that policy becomes a standard actuarial exercise against a known reference, not a bespoke negotiation against an unpublished expectation.

Research and macro desks gain a new leading indicator. Shifts in forward-curve regime — contango flattening, the migration of a humped peak from month 9 to month 6, the emergence of an inverted hump at a specific horizon — are signals about where the market expects compute supply and demand to balance. For funds running themes tied to AI infrastructure capex, this is a dataset that did not previously exist.

In each case, the common pattern is that work previously blocked by the phrase "we don't have a reference rate" can now begin. The frames above are not prescriptions. They are starting points that each desk adapts to its own book, its own risk framework, and its own counterparties.

What becomes possible in the next 18 to 36 months

The commodity-market playbook is well-documented, and there is no reason to expect AI compute to deviate from it in meaningful ways. Once a forward curve is published and referenceable, the sequence that follows is predictable.

The first cleared GPU rental swap becomes markable, then originatable, then tradeable. Neocloud bond issuances begin referencing the curve for revenue covenants, replacing bespoke modeling assumptions with market-implied trajectories. Parametric insurance products structured against forward rate movements move from pilot to origination. A term structure of implied volatility emerges from observed swap activity, which eventually supports options pricing on GPU rental rates. Basis trades develop between the forward curve and spot indices, sharpening the curve's own price discovery.

None of this is speculative. Every one of these moves has precedent in rates, oil, gas, power, and freight. The only question specific to compute is the pace — and the pace is likely to be faster than in prior commodities, because AI infrastructure is digital, cross-provider, and SKU-uniform in ways that physical commodities never were.

The implication is strategic. The financialization of AI compute is not a question of *if*. It is a question of which institutions position ahead of the sequence, and which inherit terms set by those that moved first.

The curve exists. The question is who uses it.

The AI compute market has been running without its pricing primitive. It no longer has to. The institutions that incorporate the GPU forward curve into how they underwrite, price, and hedge compute exposure will be the ones

that define the terms of this asset class for the next decade. The ones that wait for the market to be "ready" will be operating on terms set by the institutions that did not wait.

The curve is published daily. The methodology is disclosed. The data spans the full neocloud and hyperscaler landscape. What happens next depends on who is paying attention. For how the first exchange-listed contracts are taking shape on top of curves like this one, see our companion piece on GPU futures: how compute is becoming a tradable commodity.

Explore the GPU Forward Curve

Frequently Asked Questions

Is the forward rate a forecast?

No. The forward rate is a mathematical consequence of today's term-structure pricing under the no-arbitrage principle — not a prediction. It expresses the rate implied by current contracts for renting at a future instant, the same way a forward rate is derived from a yield curve.

How does the forward curve relate to the existing SDH100RT spot index?

The spot indices (A100, H100, B200) anchor the front of the curve. The forward curve extends those observable spot levels out across tenors using term-structure rates, so the two are consistent by construction: the curve begins where the spot index sits today.

Why should a no-arbitrage derivation work when GPU rental markets are fragmented?

Every commodity curve was built on a fragmented market at inception — Brent crude needed quality and basis adjustments, SOFR contended with repo segmentation, electricity still addresses zonal basis. Heterogeneity is handled through methodology transparency, not by waiting for a perfectly homogeneous market.

Which GPUs does the curve currently cover, and why not more?

Coverage starts with the most liquid, most widely rented accelerators — A100, H100, and B200 — because those are where spot data is deepest and most reproducible. The methodology generalizes to additional accelerators as their rental markets develop sufficient observable pricing.

Can the forward curve serve as a settlement reference for cleared derivatives today?

The methodology is disclosed, reproducible, and improving — the standard a settlement reference must meet. As liquidity in forward-starting contracts builds, the published curve provides exactly the market-calibrated expectation that cleared swaps and parametric products need to reference.



Written by **Carmen Li**
Founder at Silicon Data

Share this story

Subscribe to our Newsletter

Your email

Subscribe

Articles you may like

Steve Hou • Nov 16, 2026

Hopper and Ampere rents are rising, but the 30–50% is a reset, not a quarter

CoreWeave said Ampere and Hopper pricing rose 30–50% quarter over quarter. Our indices rose 2–13% in Q3, but measured from 2025 the H100 index sits inside that range.

Steve Hou • Nov 11, 2026

The Crunch Spread: GPU Rents and Token Prices Are Pulling Apart (Part 1)

GPU rents are rising and token prices are falling. We measure the crunch spread as break-even throughput: the tokens per second a GPU must serve to cover its rent.

Steve Hou • Oct 21, 2026

Dynamic Pricing Brings Price Discovery to GPU Spot Rental Markets

More than half of GPU spot price series did not change once in nine days. Nebius's new dynamic spot pricing shows what a price that moves with capacity reveals.

Silicon Data • Sep 24, 2026

B200 Rental Price Update, August 2026: The Quietest Month Yet

B200 rental index at 5.62 entering September, up 27.6% YTD: the quietest month yet, a term curve below spot, and B200 futures listing in October.

Silicon Data • Sep 15, 2026

A Compute Provider's Guide to Arbitrage: When Reserve Contracts

Silicon Data • Sep 14, 2026

The Practitioner's Guide to the Compute Future:

and GPU Futures Disagree (Part 2)

How to read a GPU reserve contract against the compute futures strip: the implied forward rate, the size of a gap, and why the two markets differ.

→

How to Hedge, Step by Step (Part 1)

How buyers and providers can lock in H100 rental costs with the new NYMEX compute futures: contract facts, a worked 50-GPU hedge, daily cash flows, basis risk.

→

Make better compute decisions today

The data intelligence platform delivering systematic visibility and institutional benchmarks across the AI economy.

Getting Started | Talk to Sales

Subscribe to our newsletter

Research and market updates on GPU compute pricing, straight to your inbox.

Your email

Subscribe

Products

- H100 Index
- H200 Index
- A100 Index
- B200 Index
- B300 Index
- MI300X Index
- SiliconMark™
- SiliconSiteIQ™
- SiliconPriceIQ™
- SiliconCarbon™
- GPU Pricing Dataset
- Token Market Pulse
- Token Pricebook
- Index & Curve Datasets
- GPU Forward Curve
- GPU Residual Value

Indices

- GPU Rental Price Indices
- H100 Index
- H200 Index
- A100 Index
- B200 Index
- B300 Index
- MI300X Index
- RAM Index
- LLM Token Index

Company

- About us
- Media
- Pricing
- Contact Us

Resources

- Blog
- Use Cases
- Events
- Market Research

Documentation

Terms of Use California Privacy Notice Privacy Notice Report a Concern Resources

© 2026 Silicon Data® is a registered trademark of Silicon Data Inc. All rights reserved.

Ask AI for a summary of Silicon Data