

Back to Research

The Basis Gap in Compute Acquisition Structures

January 27, 2026

Why Basis Risk Defines Every Procurement Decision

Every decision to procure GPUs or compute capacity in today's AI landscape is, at its core, a risk-reward trade-off - a strategic decision on how to bundle and allocate financial exposures among the parties involved: the customer / workload owner, the operator, the lessor / financier, the provider, or the broader market. In this piece, the decision-maker is the compute fleet owner - the entity responsible for ensuring GPUs are available to run workloads and absorbing the economic consequences if capacity, pricing, or hardware assumptions prove wrong. Buying, leasing, and reserving capacity each represent a different way of

packaging these risks, with no “free lunch” - you’re always choosing who ends up holding the downside when things don’t go exactly as planned.

The procurement economics hinge primarily on three key risks:

1. Spot / replacement risk - the chance that market prices spike or better hardware becomes available when the compute fleet owner is locked in to older terms, negatively affecting their unit economics due to higher refresh costs or outdated performance. This risk includes two distinct but related dynamics: (i) exposure to market pricing when capacity can’t be re-procured quickly, and (ii) exposure to hardware performance curves when new generations improve the performance-per-dollar.

2. Utilization risk (“short utilization”) - paying for capacity that goes unused. This is one of the most common hidden components - costs are fixed or committed but demand and workload timing is uncertain. They are “short utilization” not simply because lower utilization is undesirable, but because the downside from under-utilization is *asymmetric* and *nonlinear*: when utilization falls, fixed costs are spread across fewer GPU-hours and causing unit costs to rise sharply while upside from higher utilization is capped by total available capacity. For example, if the compute fleet owner commits $1,000 for 400 GPU-hours but only runs 200, their

effective cost jumps to $5 / hour from the original $2.50 / hour. In many agreements this is explicit via minimum commits or take-or-pay terms, meaning the fleet owner owes the dollars whether the workload shows up or not.

3. Residual value risk - what the GPU or cluster is worth at the end of its term. If the compute fleet owner owns hardware, it's the literal resale value; if they don't own hardware, the risk still exists economically through contract terms via optionality or locking you into a particular spec. It is dependent on timing, location, and tech cycles.

The key concept that ties these together - explaining how “locking in A” is not the same as “locking in B” - and the north star of this article, is the idea of *basis risk:*

```
Basis = (what a contract guar
```

When the gap widens, the compute fleet owners pays the difference. You can be “hedged” on paper but still wrong in practice - basis risk is why two seemingly identical deals can end up costing vastly different.

Compute fleet owners have valid, varying objectives, whether minimizing upfront capital, maximizing uptime certainty, or retaining asset control, and no single structure is inherently superior. Basis risk, however, is the common thread that

cuts across all of them, turning a theoretically sound choice into an expensive one when the contracted guarantee diverges from actual need. This risk-trading lens explains why procurement feels high stakes today, as rapid obsolescence, volatile demand, and massive capex mean small contract mismatches can explode into large P&L outcomes.

The Commitment vs. Reservation Split

Although commitment and reservation are frequently discussed in the same breath as ways to lock in compute, they serve different purposes - price and deliverability.

A commitment is a pricing contract. The capacity buyer commits to a minimum spend or usage level (e.g. 50% off spot rates for 3 years) in exchange for a lower rate. It's a hedge that reduces the buyer's unit rate, but doesn't necessarily guarantee that exact capacity will be available where or when the workload needs it. Reservation is an availability contract. The capacity buyer reserves a determined pool of capacity (specific GPUs, region, spec, etc.) so it is there when the workload needs to run. It's a hedge on deliverability, but usually at full or near-market rates with no built-in discount. The split is where the basis risk starts, as the buyer can win on one leg and still lose on the other.

Two quick examples:

1. A mid-tier customer looking to buy capacity commits to discounted H100s at $1.50 / hr vs $2.50 / hr spot (figures are purely illustrative). The price looks solid, but a regional shortage means the buyer can't actually get the GPUs during the launch window. They're forced to either delay the launch or burst onto on-demand elsewhere at $2.50 / hr - so the "discount" only applies to capacity they can't access when it matters.

2. A customer wanting to buy capacity reserves a block of capacity - say a specific region and networking tier - to guarantee availability. A few months later, the workload shifts to require a different GPU class. The reserved block is now the wrong deliverable. The customer still pays for it (or eats penalties), but to run the workload they now also have to procure new capacity in the correct GPU class / spec (and sometimes region) - effectively paying twice: once for stranded reserved capacity and again for incremental spot or new reservations.

The bottom line: commitment without reservation leaves the capacity buyer exposed to being right on price but wrong on access. Reservation without commitment leaves the buyer exposed to being right access but wrong on the workload's actual deliverable. Either way, the mismatch shows up as basis,

because what the contract guarantees is now what the workload ends up needing.

Three Structures as Synthetic Risk Positions

Buying, leasing, and reserving are three different contracts with three different failure modes. The structures look different on paper, but economically, they behave like synthetic positions: each one leaves the compute fleet owner long and someone short by key variables, and the spread between what's contracted and needed shows up as basis.

Here's a quick table to help the exposures resonate:

Structure	Who owns hardware	Long / s exposur
Buy (Own)	Compute fleet owner	Short utilizatio (fixed co uncertai demand long residual value (ov asset)
Lease	Lessor / Financier	Short utilizatio (fixed lea paymen uncertai demand neutral

		residual value exposur (asset pr risk push to lesso
Reserved Capacity	Provider / Cloud	Utilizatic risk push downstr m via tak or-pay terms; neutral / residual upside

A quick intuition on the “long” and “short” framing: this is simply operating leverage. When capacity costs are fixed or committed, higher utilization improves unit economics gradually, but lower utilization hurts disproportionately. The downside from under-utilization is larger than the upside from over-utilization. Consider a hotel with $80,000 of fixed monthly costs and 100 rooms. At 80% occupancy, costs average $100 per room-night. If occupancy rises to 100%, costs fall to $80 per room-night. But if occupancy falls to 60%, costs rise to $133 per room-night. A 20-point increase improves unit economics modestly, while a 20-point decrease worsens them much more. GPU fleets behave the same way: fixed capacity costs are spread across fewer GPU-hours when utilization falls, which is why the exposure is described as “short utilization.”

Owning hardware is the cleanest legally but the messiest economically because

the fleet owner takes the full stack. The fleet owner is short utilization as costs are fixed while demand is uncertain, and long residual value because they own the asset's future price - whether it holds or collapses. If the market reprices the hardware curve or if a new generation of hardware slashes the $ / FLOPs faster than expected, costs rise and resale values fall. Basis occurs as the deliverable the fleet owner bought - whether GPUs, interconnects, or clusters - doesn't match the evolved workload. Ownership then forces the owner to absorb the mismatch as your problem to solve.

Leasing may shift the ownership but it doesn't remove the exposures that dominate the day-to-day economics: utilization and cost of capital. The compute fleet owner remains short utilization as payments are fixed but revenue is not, so keeping busy is imperative. The primary difference is that residual value risk is pushed toward the financier / lessor, and procurement costs can inherit rate sensitivity. Even if GPU economics are solid, lease economics can worsen when the stack reprices, proving why the structure of the lease matters. Look at the triple-net (NNN) concept, for example: the lessor wants a clean yield stream backed by the asset, while the lessee absorbs operating variability (power, maintenance, hosting) and execution risk. We can see this logic in large AI infrastructure financings - e.g., Apollo's January 2026 $3.5 billion capital solution backing Valor Equity Partners $5.4 billion acquisition and triple-net lease of NVIDIA GB200 GPUs to xAI. What's important here is the allocation:

leases force the operator to carry utilization risk while the capital stack prices residual value and credit, and it creates a new basis channel if rates move or residuals underperform.

Reserved capacity changes the feel of the contract as the capacity buyer is explicitly buying delivery. This delivery tends to be defined by a region / zone and an instance / GPU class, occasionally with a performance or networking tier baked in. When the buyer reserves capacity, they lock in the shape of what is being purchased - GPU type, region, spec - so they are trading off flexibility for clarity. As a result, this creates a technology lock-in: if a workload shifts region or the “best-fit” GPU changes as the work advances, the pivot to new capacity brings friction (penalties, repricing). The basis failure here is “right on capacity, wrong on fit”, as the capacity buyer can end up paying for reserved capacity that doesn’t match the workload, while still needing to procure incremental capacity elsewhere.

No single structure solves everything; each one simply redistributes the risks in its own way. The meaningful question is which redistribution best matches the compute fleet owner’s tolerance for the basis that will inevitably appear when the hardware, workloads, or utilization diverge from what was contracted. In practice, that alignment is what separates sustainable economics from gradual erosion.

The Institutional Financing Stack

When GPU fleets scale to thousands of units across sites, procurement becomes structured finance. The workhouse tool is the special purpose vehicle (SPV): a separate legal entity created to own a defined pool of GPUs (or full cluster), borrow against that pool, and keep the debt / risk isolated from the compute fleet owner's main balance sheet.

For example, Lambda Labs structured a $500M loan in 2024 as a GPU-backed asset-backed securitization (ABS) via a dedicated SPV. The SPV held the physical Nvidia chips plus rental contracts / cash flows from cloud customers, allowing investors to fund the vehicle while Lambda used proceeds to buy more GPUs. Coreweave has taken this further with multiple delayed-draw term loans (DDTLs) through subsidiaries / SPVs (e.g., $2.6B facility closed in July 2025), collateralizing fleets to finance OpenAI and other commitments without loading their own balance sheet.

Lenders don't underwrite "a pile of GPUs" in isolation - they underwrite a complete setup: hardware resale value is paired with contracted revenue stream and supported by enforceable legal terms. The hardware provides recovery in default, but the contracts dictate reliable cash flows over time - who is paying (counterparty), how long (term), and under what pricing and cancellation terms. Enforcement mechanics are key

because the GPUs operate in live clusters - lenders need step-in rights and access if default hits.

The core basis is that collateral value tracks the volatile GPU secondary market, while revenue ties to slower AI demand, utilization, and fixed contracts. When the two drift (e.g., oversupply crashes residuals while contracts stay steady), basis risk widens.

It is valuable to reframe lender underwriting as two parallel questions:

1. How strong are the cash flows? - Counterparty quality (e.g., investment-grade tenant or major AI lab), term / cancellation (non-cancelable 3 - 5+ years, plus hell-or-high-water clauses), pricing mechanism (e.g., fixed rates), concentration (no single tenant > 50 - 70%), and deliverability milestones (e.g., full funding only after power-up and testing).

2. How real is the collateral? - Timing adds another mismatch: DDTLs commit capital upfront but draw in tranches as GPUs arrive / deploy / energize. For example, your first draw funds Gen N GPUs and you later draw funds for Gen N+1, but your customer contract is fixed for 24 months - essentially the economics of the fleet and revenue are misaligned on different clocks.

Importantly, repossession value is not simply “what is a B200 worth?” - it’s “what is a B200 worth *where it sits.*” In practice, GPUs are wired into live clusters (InfiniBand, cooling, power) -

extraction costs thousands per rack takes weeks or months, and causes downtime. This is why SPV structures obsess over hosting and access terms: control rights and hosting agreements can matter just as much as the equipment itself.

At scale, this is project finance - lending against a defined asset pool and contracted cash flows - rather than unsecured corporate debt against the computer fleet owner's balance sheet. The true risk is alignment, as to whether the collateral packaged stays aligned with the revenue stream you're reliant on.

Managing Basis at Scale

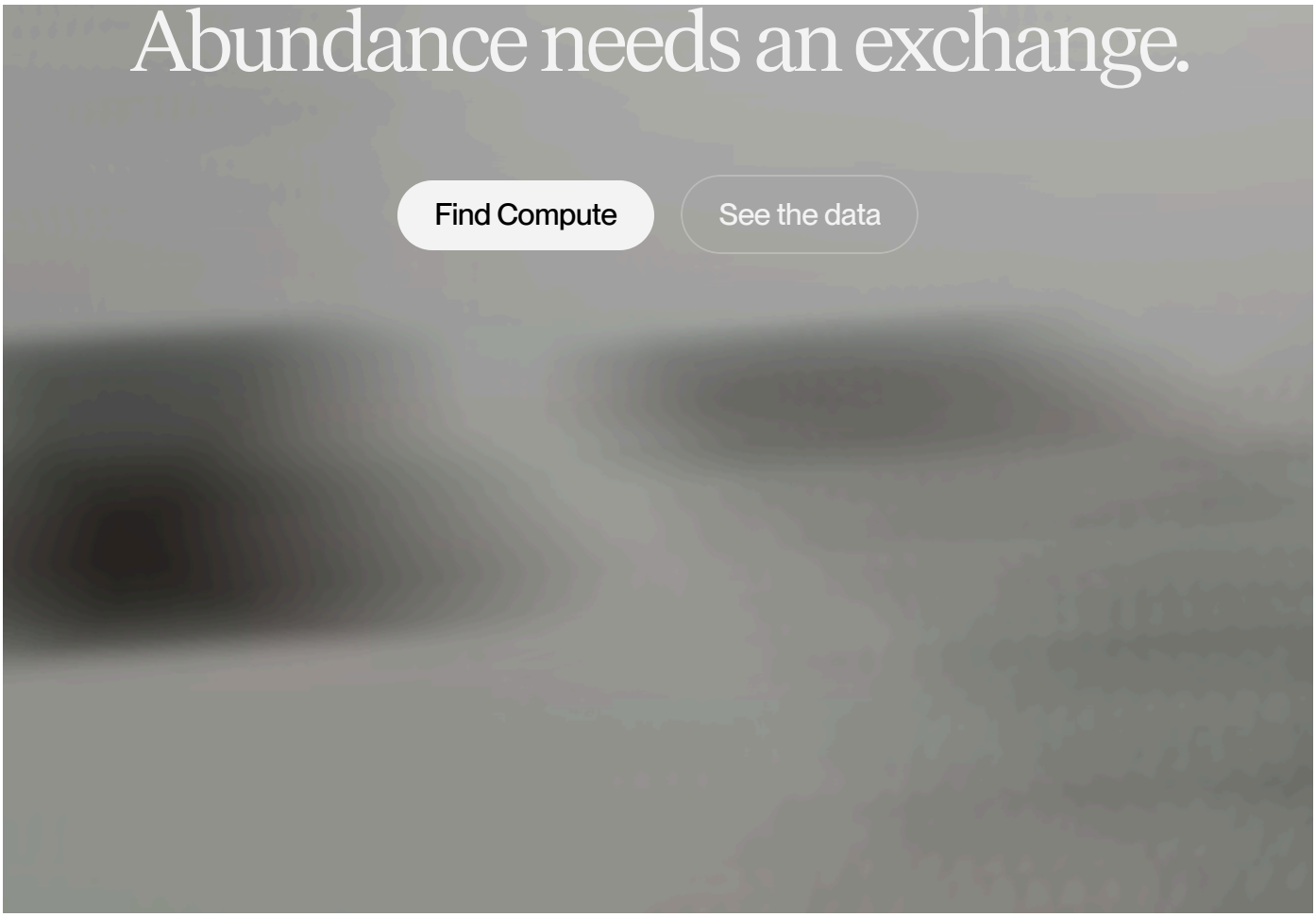
In real-world execution, basis risk isn't an abstract footnote, but rather it's the difference between a fleet that generates predictable returns and one that bleeds capital. The structures examined aren't competing solutions - they are different ways of slicing the same pie of exposures. The question is never "which one is best," but "which risks are you structurally willing to own, and which can you afford to let someone else carry?"

What makes this current moment different from prior cycles is the velocity. Generations turn over in 12 - 18 months, not 3 - 5 years. Utilization can swing 30 points in a quarter based on model efficiency or training / inference mix, and residuals can evaporate meaningful

value depending on supply. The basis gap opens faster and wider than ever, and contracts that look bulletproof on day one can become anchors in just a few months time.

Taking a step back, the compute fleet owners and financiers who will win are the ones who treat basis as the primary variable and not an afterthought. They define the deliverable with precision and map exposures, all the while building in optionality - whether through tranche-based commitments or emerging hedging tools. As compute acts more like oil, edge belongs to those who procure compute intelligently.

The future of mid-tier data center economics will be defined less by who can build the biggest fleet and more by who can manage the basis most effectively. In this game, the real risk isn't GPU scarcity - it's watching your margins evaporate when the market reprices faster than your contract can follow.



X
LinkedIn
Substack

Careers
About
Research
Docs
Trust Center
Regulatory

Privacy Policy
Terms of Use

Brand + Website by Haptic

©2026 Ornn AI, Inc. All rights reserved.

