
The Accidental Kingmaker:

How Bottleneck Power Migrates Through the AI Stack

Technology markets do not eliminate constraints. They move them. But the constraint and the bargaining power attached to it do not always move together. HBM shows the difference, and why it matters to anyone who buys, builds or invests in the AI stack.

CORE INSIGHTS

- **Scarcity is not leverage.** A constrained component gains bargaining power only when scarcity combines with a market structure suppliers can exploit and buyers who cannot wait.
- **Constraints and power move separately.** In 2026, HBM stayed constrained while power shifted within the layer, from one dominant supplier to all three. Share eroded; pricing did not.
- **We forecast the share shift correctly and the pricing erosion wrongly.** We score that miss openly and use it to refine the model.
- **The M37 Bottleneck Power Score** now separates four questions: Is there a shortage? Can suppliers exploit it? Can buyers wait? Can buyers pay?
- **The next bottleneck is a set of candidates, not a schedule.** Packaging, interconnect and power are plausible next layers, and several may bind at once.

WHERE DOES THE POWER GO?

Every complex system runs into a cascade of hard limits: compute, memory, connectivity, packaging, power, cooling. Solve one, and you expose the next. Most of these limits are simply expensive problems. Capital and engineering work around them.

Occasionally, one becomes something else. A component that looks like a procurement problem today becomes a strategic dependency tomorrow, not because the technology got harder, but because the market structure handed someone leverage.

The question is not “Where is the bottleneck?”
It is “Where does the bargaining power go, and who holds it while the bottleneck lasts?”

High Bandwidth Memory (HBM) is our evidence, not our thesis. The thesis is about how economic power moves when a technical constraint moves through a technology stack. If it holds, it should apply well beyond semiconductors, to electricity grids, data center capacity and EV supply chains. Those are tests for future articles. Here we test it against HBM.

This article follows a deliberate sequence: **observation** → **hypothesis** → **framework** → **prediction** → **outcome** → **refinement**. In 2024 we made specific predictions. Below, we score them, including the one we got wrong, and show how that miss changed the model.

HOW TO READ THE EVIDENCE LABELS

OBSERVED **Reported** results, disclosures or market data.

ANALYST ESTIMATE **Third-party** analyst estimates or projections.

M37 VIEW **Our** interpretation of the evidence.

M37 FORECAST **Our** forward-looking call, to be scored in future articles.

THE SCORECARD: WHAT WE PREDICTED, WHAT HAPPENED

In mid-2024, HBM looked different from the commodity DRAM market around it. SK Hynix held roughly half the HBM market and more than 90% of premium HBM3. Each GPU design had to be qualified with a specific memory supplier, a 12–18 month process. On that basis we made three forecasts:

- (1) HBM would become a material and growing share of GPU cost: 12–15% of bill of materials in 2024, rising to 15–20% by 2026.
- (2) SK Hynix would hold more than 40% market share through 2025, protected by qualification lead times.
- (3) Competitors would qualify HBM4 and begin eroding SK Hynix’s pricing power by late 2026.

Prediction	Evidence through September 2026	Status
① HBM becomes a material share of GPU cost	Estimated 15–20% by mid-2026; HBM prices forecast to rise further in 2027 [1]	✓ Confirmed
② SK Hynix holds >40% share through 2025	~ 50%+ through 2025; 58% in Q1 2026 and 50% in Q2 2026 [5]	✓ Confirmed
③a Competitors qualify and take share by late 2026	Samsung 33% and Micron 18% of HBM revenue in Q2 2026 [5][6]	✓ Confirmed
③b Competition erodes pricing power by late 2026	2027 blended HBM ASP forecast +121% (includes mix shift to HBM4/4E); Micron gross margin 87% [1][3]	✗ Missed

Our supply-side forecast held. Our pricing forecast did not. We split forecast ③ in two because the market split it. Competitors qualified and took share exactly as expected, yet pricing pressure intensified:

- **ANALYST ESTIMATE** TrendForce projects the blended average selling price across all HBM products to rise 121% year-over-year in 2027, with supply constrained through 2027 [1]. “Blended” matters: part of the increase reflects the shift in mix toward higher-priced HBM4 and HBM4E, so this is not a like-for-like price increase for any single product.
- **OBSERVED** Micron reported fiscal Q4 2026 revenue of \$54.2 billion at an 87.0% non-GAAP gross margin and says it has contracted “the vast majority” of its calendar 2027 HBM supply “with significant price increases year-over-year” [3][4].
- **OBSERVED** Spot HBM3E reportedly traded at four to five times long-term contract prices in late August [2].
- **OBSERVED** Nvidia’s CFO said in August that memory price increases had “exceeded our prior expectations” and were “headed even higher into next year.” SK Hynix’s CEO reportedly does not expect supply and demand to balance before 2030 [7].

None of these measures captures bargaining power directly. Taken together, our evidence indicates that **supplier pricing power persisted and may have strengthened.**

We publish the miss because it is the most useful data point, we have. It is what forced the refinement at the center of this article.

THREE LEVELS OF CONSTRAINT

Most supply-chain analysis stops at “scarce, therefore powerful.” We separate three levels:

Level	What it means	The test
1. Physical constraint	There is not enough capacity.	Can buyers get what they need, when they need it?
2. Economic bottleneck	Suppliers can extract economic rent from the shortage.	Are prices and margins rising faster than costs?
3. Strategic bottleneck	The constraint changes what the customer can do.	Is it reshaping designs, roadmaps or system pricing?

A component can sit at level 1 without reaching level 2, if buyers have alternatives or can wait. That is the core distinction: **scarcity alone does not create leverage.** But HBM in 2026 adds a qualification: a severe enough shortage can produce leverage even as market structure weakens.

The mechanism matters. Structure normally creates leverage because it removes the buyer’s best alternative, another supplier. When every qualified supplier is sold out, that alternative disappears anyway. The buyer’s real choices become waiting, redesigning, or paying whatever price clears the market. If waiting is costly and redesign takes a product cycle, the clearing price rises for every supplier at once. Under those conditions, **extreme scarcity can stand in for structure.** That is a significant change to our model, and Test 5 below exists to keep checking it.

M37 VIEW HBM reached all three levels in 2024 and remains at all three in 2026. Memory cost now flows directly into system prices: Nvidia has reportedly told customers AI server prices will rise more than 15% from early 2027 [8], and TrendForce reports GPU and ASIC makers considering a move from 12-high to 8-high stacks, reducing HBM per chip, in response [1].

WHAT HBM REVEALED: POWER MOVED, THE CONSTRAINT DID NOT

Our 2024 model assumed that once competitors qualified, the constraint would ease and pricing power would fade. What happened was different. The constraint held, but the power attached to it moved within the layer.

OBSERVED Counterpoint estimates SK Hynix's HBM revenue share fell from 58% in Q1 2026 to 50% in Q2, while Samsung rose from 21% to 33% and Micron held 18% [5]. Samsung had begun shipping HBM4 to Nvidia in February 2026 [6]. *(Share estimates vary by revenue versus bits and by product generation.)*

The kingmaker lost share. The kingdom grew. Power shifted from one supplier to the category.

A caution: the evidence does not show that all three suppliers captured equal economic rent. It shows that **supplier concentration fell without relieving category-wide scarcity**.

Why we missed it

We modeled competition as a contest for share, not capacity. New suppliers added names but not enough bits. When every qualified supplier is sold out, another entrant changes who collects the rent, not how much rent there is.

We underestimated how far HBM would squeeze adjacent memory. HBM uses considerably more wafer capacity per bit than conventional DRAM. TrendForce data, as reported, show DRAM contract prices rising roughly 90–95% quarter-over-quarter in Q1 2026 and 58–63% in Q2 [10]. One caution on causation: this was not only the system pushing the constraint sideways. Suppliers *chose* to prioritize higher-margin HBM. We cannot yet separate structural spillover from deliberate allocation.

We treated financial pressure as if it reduced urgency. The two are separate: a buyer can be urgent and financially constrained at the same time. Buyers stayed urgent and protected their capacity to pay by borrowing. S&P Global projects that only one of the six largest AI spenders will generate positive free cash flow in 2027, yet they are responding by borrowing, not slowing [11]. It puts their combined 2026 capex near \$870 billion [11], and Goldman Sachs projects hyperscaler AI capex rising about 50% to \$1.2 trillion in 2027 [12].

We misjudged custom accelerators. We expected them to diversify demand away from the bottleneck. Instead, they added a second demand pool for the same constrained bits; ASIC demand was one of the drivers TrendForce cited for HBM3E price increases in 2026 [9].

THE REFINEMENT

Constraint migration and power migration are separate events. A constraint can stay put while the power attached to it moves, between suppliers within a layer, or sideways into adjacent markets.

Tracking only where the shortage is will miss where the leverage is.

THE MECHANISM: CONSTRAINT × STRUCTURE × URGENCY

Our working model is conceptual, not an equation. The pillar scores below are not units, and multiplying them would not produce a meaningful number. The point is the shape of the relationship:

M37 WORKING MODEL

Bottleneck power increases with Constraint × Market Structure × Customer Urgency, enabled by Capital Capacity.

Pillar	The question	What drives it
Constraint	Is there a shortage?	Supply headroom: how long capacity is sold out; spot versus contract prices.
Market Structure	Can suppliers exploit it?	Supplier concentration, qualification time, replication cost, switching cost. These change over years.
Customer Urgency	Can buyers wait?	Expected returns, lack of substitutes, deployment deadlines. These change over quarters.
Capital Capacity (enabler)	Can buyers pay?	Cash flow and access to debt markets. Urgency without capital cannot sustain a premium.

The relationship is multiplicative, not additive. If any pillar is near zero, there is no bottleneck power: a shortage with many interchangeable suppliers is just a shortage, and a concentrated market with spare capacity is just a concentrated market.

It also means weakness in one pillar can be outweighed by strength in another. That is what the 2026 HBM data suggest. Market Structure weakened as SK Hynix lost its dominant position. Constraint intensified as supply sold out into 2027. Urgency stayed high, and Capital Capacity, though strained, held. The net effect: supplier pricing power persisted and appears to have strengthened.

M37 VIEW Two pressure points are worth watching. The first is Capital Capacity: S&P Global projects only one of the six largest AI spenders will have positive free cash flow in 2027 [11]. So far, they have borrowed rather than cut back; if debt markets tighten, this is where the model would show it first. The second is substitution: buyers have not regained leverage through negotiation, but they are starting to through design, considering 8-high instead of 12-high stacks [1]. When customers cannot negotiate around a bottleneck, they engineer around it.

THE M37 BOTTLENECK POWER SCORE

A working hypothesis for measuring where leverage sits

To make the model usable, we score each pillar. This is a hypothesis and a measurement framework, not a calibrated index. Scores are directional; they become more meaningful as we score more components over time and check them against outcomes.

Scoring rubric (0–10 per dimension)

Dimension	Low (0–3)	Medium (4–6)	High (7–10)
CONSTRAINT			
Supply headroom (tightness)	Spare capacity; spot at or below contract	Roughly balanced; sold out ≤2 quarters	Sold out 12+ months; spot far above contract
MARKET STRUCTURE			
Supplier concentration	4+ suppliers	2–3 suppliers	1 dominant
Qualification time	<6 months	6–12 months	12–18+ months
Replication cost	<\$500M	\$500M–\$2B	\$2B+
Switching cost	<\$10M per design	\$10–100M	\$100M+ per design
CUSTOMER URGENCY			
ROI urgency	Elastic; can wait	Moderate	Must deploy now
CAPITAL CAPACITY (ENABLER)			
Capital availability	Constrained	Moderate	Abundant

Pillar scores

Pillar	HBM 2024–25	HBM 2026–27	CoWoS 2026–27
Constraint	9	10	7
Market Structure (avg. of 4)	9.0	8.0	7.8
Customer Urgency	8	9	8
Capital Capacity (enabler)	9	8	8

Market Structure dimension scores (concentration / qualification / replication / switching): HBM 2024–25: 9/9/9/9. HBM 2026–27: 6/9/9/8. CoWoS 2026–27: 8/8/9/6. All scores are M37 judgments.

Read across the HBM columns and the 2026 story is visible at a glance: Structure fell, Constraint rose, Urgency rose, and Capital Capacity eased but stayed high. A single averaged score would have hidden that.

The bottleneck matrix

Plot Market Structure against Customer Urgency, with Constraint as a gate and Capital Capacity as an enabler. Without a real shortage (Constraint below about 5), even a strategic-looking position is dormant. Without the capital to pay, urgency cannot sustain a premium.

	Low urgency	High urgency
Weak structure	Commodity <i>Buyers hold the leverage</i>	Temporary shortage <i>Prices spike, then normalize as supply responds</i>
Strong structure	Latent bottleneck <i>Leverage waits for demand urgency</i>	Strategic bottleneck <i>HBM 2024–25 · HBM 2026–27 · CoWoS 2026–27</i>

All three cases sit in the strategic quadrant, but for different reasons. HBM 2024–25 got there through structure. HBM 2026–27 stays there through constraint. CoWoS combines strong structure with high urgency, but its constraint score is lower because capacity is expanding fast. Whether that is enough to create pricing power is the test for Article 3.

WHERE POWER MOVES NEXT: CANDIDATE MIGRATION POINTS

We do not offer a schedule of future bottlenecks. We offer candidates: layers that show the conditions for bottleneck power and that we will score and test. The first “migration” is already visible, and it went sideways rather than upstream.

Already observed: conventional memory

OBSERVED HBM’s claim on DRAM capacity turned commodity memory into a constrained market in 2026, with price effects reaching servers, PCs and consumer devices [10]. Whether this was migration or supplier allocation (or both) is an open question.

Candidate 1: Advanced packaging (CoWoS)

ANALYST ESTIMATE TSMC, the dominant source of CoWoS capacity, grew from roughly 35,000 wafers per month in late 2024 [16] to an estimated 130,000 by end-2026, and reportedly plans to reach about 260,000 by end-2028 [13]. Estimated 2026 demand is near 1.0 million wafers, with Nvidia securing roughly 60% of allocation [14][15]. The supply-demand gap was projected to narrow from about 20% to about 10% by end-2026 [17].

ANALYST ESTIMATE A second source is emerging: Intel’s EMIB-T packaging is projected at 15,000–20,000 CoWoS-equivalent wafers per month in 2027 and 40,000–45,000 in 2028 [13].

M37 VIEW The HBM lesson sharpens the question for Article 3: will a second source erode TSMC’s pricing power, or, as with HBM, merely redistribute share while capacity stays sold out?

Candidate 2: Chiplet interconnect

M37 VIEW As monolithic designs approach cost and power limits, chiplet interconnects carry HBM-like barriers: difficult qualification, concentrated manufacturing and design-embedded switching costs.

Candidate 3: Power delivery

M37 VIEW Power availability at the data center and grid level is a physical constraint today. Whether it becomes an economic bottleneck depends on who controls access to it, which differs sharply by region.

Not a chain, possibly a crowd

A simple migration story says Bottleneck A is solved, then Bottleneck B appears. The 2026 evidence suggests something messier: HBM, conventional DRAM and advanced packaging were all constrained at the same time. If several layers bind at once, power may fragment across them rather than move cleanly from one to the next. We may be looking at **constraint competition**, not just constraint migration. We add this as an explicit test below.

HBM outlook

M37 FORECAST HBM pricing power persists through 2027. The earliest plausible erosion window is 2028, and only if new capacity outpaces demand. Micron’s Singapore HBM packaging facility begins initial output in early 2027 [4].

Signals that erosion is starting:

- The spot-to-contract premium narrowing.
- Memory contract price increases continuing to decelerate (DRAM increases slowed to roughly 13–18% quarter-over-quarter in Q3 2026 [10]).
- Wider adoption of lower-HBM chip configurations.
- Hyperscaler capex cuts or stress in the debt markets now funding AI buildouts [11].

WHAT WOULD FALSIFY THIS?

A useful hypothesis must be able to fail. These are our tests, including two added after the HBM outcome.

Test	What would falsify or force revision	Status (Oct 2026)
1. Constraint without power	CoWoS stays severely constrained through 2027, but suppliers gain no meaningful pricing power.	Open: Article 3
2. Substitutes accelerate	Buyers engineer around a bottleneck faster than qualification barriers can protect it.	Early signals: less HBM per chip; Intel EMIB-T
3. Capacity outruns demand	Capacity consistently arrives before scarcity creates leverage.	Not met in HBM
4. Capital overwhelms scarcity	Abundant capital, not market structure, explains supplier power.	Inconclusive
5. Share without price (new)	A dominant supplier loses significant share while category prices rise.	Met in HBM: model revised
6. Constraint competition (new)	Several layers bind at once and power fragments across them rather than migrating.	Early evidence in 2026; tracking

Test 5 did not refute the model; it refined it. Market structure explains *who* holds power. Constraint explains *how much* power there is. Test 6 could change the model more fundamentally, and we will report on it either way.

EXECUTIVE PLAYBOOK: FIVE QUESTIONS

1. What can stop us?

Map every dependency in your architecture and supply chain that would delay a product or deployment if it became unavailable.

2. Who has leverage over us?

Score each dependency on Constraint, Market Structure, Customer Urgency and Capital Capacity. High scores across the board signal a strategic bottleneck.

3. Is scarcity creating bargaining power?

Separate a shortage from a bottleneck. Is anyone holding spare capacity? Are you locked in? Can you wait? And remember: a qualified second source is not necessarily a cheaper one while the category is sold out.

4. What becomes scarce if this constraint disappears, or alongside it?

Look upstream and sideways. HBM tightened conventional DRAM before packaging became the headline constraint.

5. What should we qualify, redesign, contract or invest in now?

By the time pricing power becomes obvious, the companies paying attention have already started solving the bottleneck. Qualify alternatives early, design for flexibility, and secure long-term capacity where headroom is scarce.

THE WORKING HYPOTHESIS

Our model has evolved. Version 1 said constraints migrate. Version 2 said constraints migrate, but power does not necessarily migrate with them. Where the evidence now points:

Constraint, market structure and urgency interact. When they diverge, physical scarcity and economic power can move in different directions, and sometimes several constraints compete at once.

Tracking only where the shortage is will miss where the leverage is.

Article 1 argued that there is hidden value beneath obvious technology trends. This article argues that there is hidden leverage beneath obvious technology bottlenecks. Article 3 asks whether the framework can predict where economic power emerges next.

NEXT IN THE SERIES

Article 3: **“The Packaging Tax.”** We apply the Bottleneck Power Score to advanced packaging and run Tests 1, 2 and 6: does CoWoS scarcity create pricing power, or just another constrained layer without economic leverage; does Intel’s entry erode it or merely redistribute share, and is packaging one bottleneck among several binding at once?

ABOUT THE AUTHOR

Joe Batista advises boards, C-suite executives, and go-to-market teams on technology value creation, growth strategies, and the opportunities hidden beneath major technology shifts. He has spent more than 40 years in the technology industry, guiding companies on growth and technology decisions, spanning the minicomputer era at Digital Equipment Corporation (DEC) through client-server, cloud, virtualization, and now AI, with executive experience at Dell Technologies and Hewlett Packard Enterprise.

Through M37 Advisory, Joe helps leadership teams recognize recurring patterns, uncover second-order effects, find the opportunities others overlook, and identify the strategy worth betting on. His work is grounded in a simple belief: **What if the biggest obstacle to your growth isn’t the competition, but the way you think?**

“Most executives have been bombarded with ‘theories’ about what might work. Based on our research, Joe’s extensive playbook is rooted in creating alignment of interest and business outcomes that leaders should take notice of.”

— Jim Dickie, Research Fellow & Sales Mastery Founder, CSO Insights

Author Contact

Email: Chief.Creatologist@M37-Advisory.com

Website: <https://www.m37-advisory.com>

Idea Vault: <https://www.m37-advisory.com/idea-vault>

Keywords

Constraint Migration | Power Migration | Bottleneck Power | Constraint Competition | HBM | SK Hynix | Samsung | Micron | NVIDIA | DRAM | Advanced Packaging | CoWoS | Intel EMIB | Chiplets | AI Infrastructure | Semiconductor Economics

SOURCES & METHODOLOGY

Forecasts date from mid-2024; outcomes run through September 2026; the outlook is as of October 1, 2026. Claims are labeled as observed, analyst estimate, M37 view or M37 forecast, and cited by number. Figures from secondary reporting are described as “reported.” HBM share estimates are definition-sensitive.

- [1] TrendForce. “HBM Supply Constraints Persist, 2027 Price Outlook Revised Upward with Blended ASP Forecast to Rise 121% YoY.” Press release, September 29, 2026.
- [2] Seoul Economic Daily. “Spot HBM Prices Hit 5 Times Contract Levels.” August 31, 2026.
- [3] Micron Technology. “Micron Technology, Inc. Reports Record Fiscal Fourth-Quarter and Full-Year 2026 Results.” September 30, 2026.
- [4] Micron Technology. Fiscal Q4 2026 earnings call (transcript via Investing.com). September 30, 2026.

- [5] Counterpoint Research, via Seoul Economic Daily. "Samsung Doubles HBM Market Share to 33%, Narrowing Gap With SK hynix." September 3, 2026.
- [6] TrendForce. "Samsung HBM4 Reportedly to Ship First After Lunar New Year." February 9, 2026.
- [7] Nvidia Q2 FY2027 earnings call (August 26, 2026) and SK Hynix CEO remarks, as reported by KuCoin, "Nvidia CFO Says Memory Prices Will Rise Further as SK Hynix Sees HBM Balance Only by 2030." September 22, 2026.
- [8] The Next Web. "Nvidia AI Server Prices Are Rising More Than 15% From Early Next Year." August 22, 2026.
- [9] TrendForce. "Samsung, SK hynix Reportedly Plan ~20% HBM3E Price Hike for 2026 as NVIDIA H200, ASIC Demand Rises." December 24, 2025.
- [10] Tech Insider. "GPU Relief Pushed to 2028 as DRAM Jumps 95%." September 17, 2026 (citing TrendForce quarterly DRAM contract price data).
- [11] The Motley Fool, citing S&P Global. "6 Hyperscalers Driving the AI Revolution Are Projected to Spend \$1.3 Trillion on Capex in 2027." September 6, 2026.
- [12] Bloomberg. "Goldman Sees Hyperscaler AI Capex Rising 50% to \$1.2 Trillion." September 25, 2026.
- [13] TrendForce. "TSMC Reportedly Targets 22% 2nm, 16%+ 3nm Capacity Boost by Mid-2027; CoWoS to Double by 2028." September 14, 2026.
- [14] Silicon Analysts. "TSMC Foundry Allocation 2026: CoWoS Sold Out, 2nm Booked, ~1M Wafer Demand." June 16, 2026.
- [15] Astute Group. "Advanced Packaging Demand Soars: Nvidia Secures 60% of CoWoS Capacity." August 4, 2025.
- [16] SemiWiki Forum. "CoWoS Capacity Set to Skyrocket by 2026." December 2024.
- [17] TrendForce. "TSMC CoWoS Supply-Demand Gap Reportedly Seen Narrowing from 20% to 10% by End-2026." June 15, 2026.

To subscribe to the series, complete the form [LINK](#)