Memory, Not GPUs, Now Sets the Price of AI

On 26 August, Nvidia told investors its gross margin will fall from 75% to a trough of 71–72% because memory chips cost more than it expected. Three days earlier it had warned its largest customers that AI server prices rise more than 15% from early 2027. Memory now accounts for roughly 62% of the bill of materials in a next-generation rack, up from 53% a generation ago. For the first time since the buildout began, the unit cost of AI compute is going up — and the scarce input is no longer the GPU.

AUTHOR

AT

Alicante Tech Vibes *Research Labs*

~8 MIN READ · ALL FIGURES SOURCED & DATED

62%

MEMORY SHARE OF VERA RUBIN RACK BILL OF MATERIALS, VS 53% FOR GB300 · GOLDMAN SACHS, 10 AUG 2026

71–72%

NVIDIA'S GUIDED Q4 FY27 GROSS-MARGIN TROUGH, DOWN FROM 75.0% IN Q2 · 26 AUG 2026

>15%

AI SERVER PRICE INCREASE NOTIFIED TO MAJOR CUSTOMERS, EFFECTIVE EARLY 2027 · BLOOMBERG, 23 AUG 2026

+435%

YEAR-ON-YEAR RISE IN MEMORY SPEND PER RACK, TO ~$2M OF A ~$7.8M SYSTEM · 10 AUG 2026

13–18%

FORECAST QOQ RISE IN SERVER DRAM CONTRACT PRICES IN 3Q26, AFTER +90–95% IN 1Q26 · TRENDFORCE

$0.75

GEMINI 3.7 FLASH INPUT PRICE PER MILLION TOKENS — HALF ITS PREDECESSOR · GOOGLE, 13 AUG 2026

I. The bill arrives at the most profitable company in the buildout

Nvidia's second-quarter results on 26 August were, on the surface, the best in its history: revenue of $96.2 billion, up 106% year on year, with data-centre revenue of $89.0 billion. The number that moved the story was smaller. GAAP gross margin came in at 75.0%, the company guided the next quarter to 74.0%, and on the call CFO Colette Kress went further — margins bottom "in Q4 in the 71 to 72% range before settling at 72 to 73% in fiscal year '28". The cause she named was memory.

Three days earlier, Bloomberg reported that Nvidia had notified its largest buyers of price increases of more than 15% on Grace Blackwell and Vera Rubin systems shipping from early 2027, sized according to product generation and memory configuration. **Taken together these are the same event seen from both sides of the income statement: Nvidia is absorbing part of a memory cost shock and passing the rest downstream.** Roughly four points of gross margin at a $400-billion-a-year revenue run-rate is real money, and the company chose to disclose the trough rather than discover it quarter by quarter.

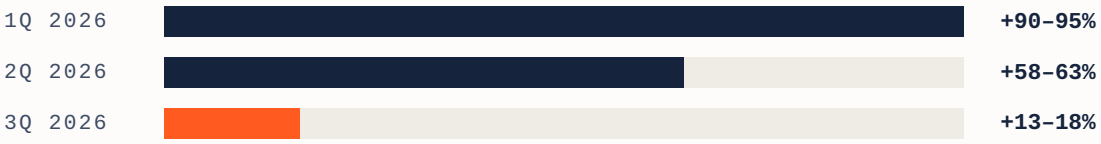
II. The chokepoint moved, and nobody re-priced it

For three years the binding constraint on AI was advanced packaging and GPU die supply — a bottleneck Nvidia controlled and monetised. That is no longer where the scarcity sits. Goldman Sachs' teardown of the Vera Rubin platform, published 10 August, puts memory at about 62% of total material cost, against 53% for the GB300 generation, with memory content alone approaching $2 million of a roughly $7.8 million rack and memory spend per system up more than 435% year on year.

When one input crosses 60% of a system's bill of materials, pricing power migrates to whoever supplies it. That is now three firms: Samsung, SK hynix and Micron. The shift is visible in Nvidia's own engineering. Analysts at GF Securities report that Nvidia has halved SOCAMM module capacity from 192GB to 96GB, cutting CPU-attached memory per rack from 54–55TB to about 28TB, on Bernstein estimates that HBM4 could reach $53 per gigabyte in 2027. A chip

designer trimming memory out of its flagship to protect a cost target is not an optimisation. It is a supplier setting the specification.

FIG. 1: DRAM CONTRACT PRICES — THE SHOCK, THEN THE PLATEAU (QOQ CHANGE)



BARS SCALED TO THE 1Q26 MIDPOINT (92.5% = FULL WIDTH). NOTE THESE ARE QUARTER-ON-QUARTER INCREASES STACKED ON EACH OTHER, NOT A RETURN TO EARLIER LEVELS. SOURCES: TRENDFORCE, 9 JULY 2026; TOM'S HARDWARE, 23 AUGUST 2026.

III. The honest read: a real constraint inside a famous cycle

The physical shortage is well documented. TrendForce expects server DRAM contract prices to rise 13–18% quarter on quarter in 3Q26, with RDIMM bit supply growing only 15–20% year on year — materially slower than server CPU shipments, which is why the firm expects tightness to persist into 2027. Buyers are already rationing: TrendForce observes customers migrating from 96GB and 128GB modules down to 32GB and 64GB. Consumer markets show the pass-through unfiltered, with a 32GB DDR5-6000 kit at around $392 in August against $110–140 a year earlier.

Three caveats deserve equal weight. First, the rate of increase is decelerating sharply — Fig. 1 is a plateau forming, not an accelerating spiral, and TrendForce notes consumer buyers are "reaching their affordability limit", which caps how far prices can run. Second, the pain is unevenly distributed: several large US cloud providers hold long-term agreements that contractually bar suppliers from raising their prices, so the increases land on everyone else. Third, memory is the most reliably cyclical business in semiconductors. Deloitte estimates combined capex at Micron, Samsung and SK hynix rising nearly 340% between 2024 and 2027. Every prior cycle ended the same way — new fabs land, supply floods, prices collapse. The counter-argument is that cleanroom, power and water constraints push meaningful new capacity to 2028 at the earliest.

“The magnitude of the price increase has exceeded our prior expectations and are headed even higher into next year.”

COLETTE KRESS, NVIDIA CFO, ON MEMORY COSTS — Q2 FY2027 EARNINGS CALL, 26 AUGUST 2026. SHE FRAMED THE SCARCITY AS A SYMPTOM OF THE SAME DEMAND SURGE DRIVING NVIDIA'S OWN GROWTH, AND GUIDED FISCAL 2028 REVENUE GROWTH OF ~70% AS A SUPPLY-CONSTRAINED NUMBER.

IV. Two curves are now moving in opposite directions

The same fortnight that hardware got more expensive, intelligence got cheaper. Google released Gemini 3.7 Flash on 13 August at $0.75 per million input tokens and $3.75 per million output tokens — roughly half its predecessor. This is not a contradiction; it is the central fact. Epoch AI's measurement of inference pricing found declines of between 9x and 900x per year at fixed capability, varying by task, driven by distillation, sparsity and better serving stacks — with the explicit caveat that the fastest of those drops may not persist.

Algorithmic deflation and hardware inflation are now running simultaneously, and the gap between them is being financed rather than earned. Alphabet raised 2026 capital spending to $195–205 billion from $180–190 billion. So long as capital is abundant, a token price can fall while the asset producing it costs more, because the shortfall lands in depreciation schedules rather than in the price list. That works until the cost of capital changes or the efficiency curve flattens — and one of those two is now measurably harder to sustain.

V. Who actually pays

The incidence is the story. Hyperscalers with long-term memory agreements and in-house accelerators are largely shielded; Nvidia's >15% increase reaches them through server makers but their memory input is pre-priced. The exposed parties are the ones without that cover: GPU neoclouds reselling capacity on thin spreads, enterprises buying racks outright, sovereign compute programmes with fixed budgets in national currency, and any AI company whose gross margin assumed hardware costs falling 20–30% per generation. For those buyers, a 15% price increase on a system whose memory content has been cut is a double hit — more expensive, and less capable per rack for memory-bound workloads such as long-context serving and large KV caches.

The second-order effect is a divergence between capex measured in dollars and capacity measured in useful compute. When hyperscalers report record

spending this autumn, the meaningful question is no longer how large the number is but how many accelerators and megawatts it buys. Rising input prices mean the same headline capex delivers less deployed capability — a distinction that flatters the investment narrative while quietly slowing the capability roadmap.

FIG. 2: MEMORY'S SHARE OF THE RACK BILL OF MATERIALS, BY NVIDIA PLATFORM GENERATION

BARS SCALED TO THE VERA RUBIN FIGURE (62% = FULL WIDTH). VERA RUBIN RACK BOM ESTIMATED AT ~$7.8M, OF WHICH MEMORY APPROACHES $2M. SOURCE: GOLDMAN SACHS ANALYSIS REPORTED 10 AUGUST 2026.

VI. The precedent cuts both ways

Memory has run this loop before. Demand surges, prices spike, everyone builds fabs, supply overshoots, prices collapse — Samsung, SK hynix and Micron all posted losses in 2022–23 after pandemic demand faded. That history is the single strongest reason to treat the current shortage as temporary. What makes this cycle arguably different is the shape of the demand: AI memory demand is contracted years forward, sits inside multi-year infrastructure commitments, and comes from buyers whose alternative to paying is not buying less but building less. SK hynix has earmarked ₩54.3 trillion (about $38.3 billion) through 2031 against that demand. The honest position is that both readings are live: the constraint is real for the next four to six quarters, and the cycle has never once failed to turn.

VII. How it plays out: three scenarios

Probabilities below are analytical judgment, not measured frequencies, over a horizon to end-2028.

1 · Pass-through ~55%

2027–2028 · GRINDING

Memory stays tight through 2027, Nvidia's margin troughs near 72% and recovers as pricing actions land, and the cost increase propagates to non-LTA buyers. Token prices keep falling on algorithmic gains, funded by capital rather than by cost. Watch for: Nvidia gross margin landing within 50bp of guidance in the next two prints; neocloud gross margins compressing in Q4 2026 disclosures; further downward revisions to memory content per rack.

2 · Hard constraint ~20%

2027 · FAST

Supply fails to arrive, HBM4 pricing overshoots Bernstein's $53/GB, and the shortage becomes the governing variable of the buildout. Capex budgets hold in dollars but shrink in delivered compute; frontier training and long-context serving roadmaps slip. Watch for: 4Q26 server DRAM contract forecasts re-accelerating above 20% QoQ; further memory-capacity cuts announced on next-generation platforms; hyperscalers guiding capacity targets down while holding spend flat.

3 · Cycle turns ~25%

LATE 2027–2028 · ABRUPT

The 340% capex surge lands, an AI spending pause hits demand before capacity is absorbed, and prices fall as fast as they rose. Nvidia margins return above 75%, the >15% price increase is quietly rescinded or discounted, and the episode reads in hindsight as a classic memory squeeze. Watch for: DRAM contract prices flat or negative in any quarter through 2027; memory-maker inventory days rising; any hyperscaler trimming capex guidance.

WHAT WOULD FALSIFY THIS THESIS

The claim here is that memory is a durable structural constraint that has inverted AI's hardware cost curve, not a two-quarter blip. Three observations would break it. If Nvidia's Q3 FY2027 gross margin (due late November 2026) prints at or above 75% and the Q4 trough guidance is revised upward, the cost shock is smaller than management signalled. If TrendForce's 4Q26 contract-price forecast comes in flat or negative, the plateau visible in Fig. 1 is already a peak. And if hyperscaler Q3 disclosures show accelerator deployments tracking capex dollars one-for-one, the pass-through is being absorbed without any loss of delivered capacity.

The strongest counter-case is simply the historical base rate: memory has never sustained a supply squeeze through a capex cycle of this magnitude, and combined capex at the three suppliers is set to rise nearly 340% between 2024 and 2027 (Deloitte, reported 7 August 2026). A reader who weights that base rate heavily should read this brief as describing a severe but temporary transfer of margin, not a permanent repricing.

WHAT TO WATCH, IN ORDER

1. **Nvidia Q3 FY2027 results, late November 2026** — whether gross margin lands at the guided 74.0% and whether the 71–72% Q4 trough holds. This is the cleanest single read on pass-through.
2. **TrendForce 4Q26 contract-price forecast, October 2026** — a reading below 10% QoQ confirms the plateau; above 20% signals the hard-constraint path.
3. **Samsung and SK hynix Q3 2026 results, late October 2026** — capex revisions and 2027 HBM4 allocation disclosures indicate when supply actually arrives.
4. **Hyperscaler Q3 2026 earnings, late October 2026** — look past the capex headline for deployed megawatts or accelerator counts; a widening gap between dollars and units is the diagnostic.
5. **Frontier-model pricing through Q4 2026** — whether the next Flash-class release repeats the ~50% cut of 13 August. A slowing cadence would be the first sign hardware inflation is reaching the price list.

SOURCES

1. NVIDIA, “Financial Results for Second Quarter Fiscal 2027,” 26 August 2026 · globenewswire.com
2. AlphaStreet, “NVIDIA Corporation (NVDA) Q2 2027 Earnings Call Transcript,” 26 August 2026 · news.alphastreet.com
3. The Herald Business (reporting Bloomberg), “Even Nvidia can't absorb memory costs — AI server prices to rise more than 15%,” 23 August 2026 · mbiz.heraldcorp.com
4. Tom's Hardware, “Nvidia reportedly warns biggest customers of 15% price hikes on AI servers,” 23 August 2026 · tomshardware.com
5. Crypto Briefing (reporting Goldman Sachs), “Nvidia's Vera Rubin superchip memory costs estimated at 62% of materials bill,” 10 August 2026 · cryptobriefing.com
6. DigiTimes, “Memory drives 62% of Nvidia Vera Rubin's cost — SOCAMM2, not HBM4, leads the bill,” 10 August 2026 · digitimes.com
7. Wccftech (reporting GF Securities and Bernstein), “NVIDIA Trims Vera Rubin Memory as HBM4 Prices Threaten to Eat 29% of Every Rack's Cost,” 26 July 2026 · wccftech.com
8. TrendForce, “Long-Term Agreements Cap Price Increases; Server DRAM Contract Prices Expected to Rise 13–18% QoQ in 3Q26,” 9 July 2026 · trendforce.com
9. TrendForce, “AI Server Demand Continues to Support Memory Prices in 3Q26, but Gains Moderate,” 3 July 2026 · trendforce.com
10. Invezz (citing Deloitte), “What Samsung, SK Hynix and Micron's capex mean for the future of the memory trade?” 7 August 2026 · invezz.com
11. InfoWorld, “Google cuts Gemini 3.7 Flash prices as enterprise AI economics diverge,” 14 August 2026 · infoworld.com
12. Epoch AI, “LLM inference prices have fallen rapidly but unequally across tasks,” 12 March 2025 · epoch.ai
13. Data Center Dynamics, “Google increases 2026 capex to $195-205bn,” 22 July 2026 · datacenterdynamics.com

METHOD NOTE: FIGURES ARE ATTRIBUTED AND DATED INLINE; WHERE INDEPENDENT SOURCES DISAGREE, THE DISAGREEMENT IS REPORTED RATHER THAN AVERAGED. BILL-OF-MATERIALS AND HBM4 PRICING FIGURES ARE THIRD-PARTY ANALYST ESTIMATES, NOT COMPANY DISCLOSURES, AND ARE LABELLED AS SUCH. SCENARIO PROBABILITIES ARE ANALYTICAL JUDGMENT.

Built by the Alicante Tech Vibes community, a tope. 2026-08-27