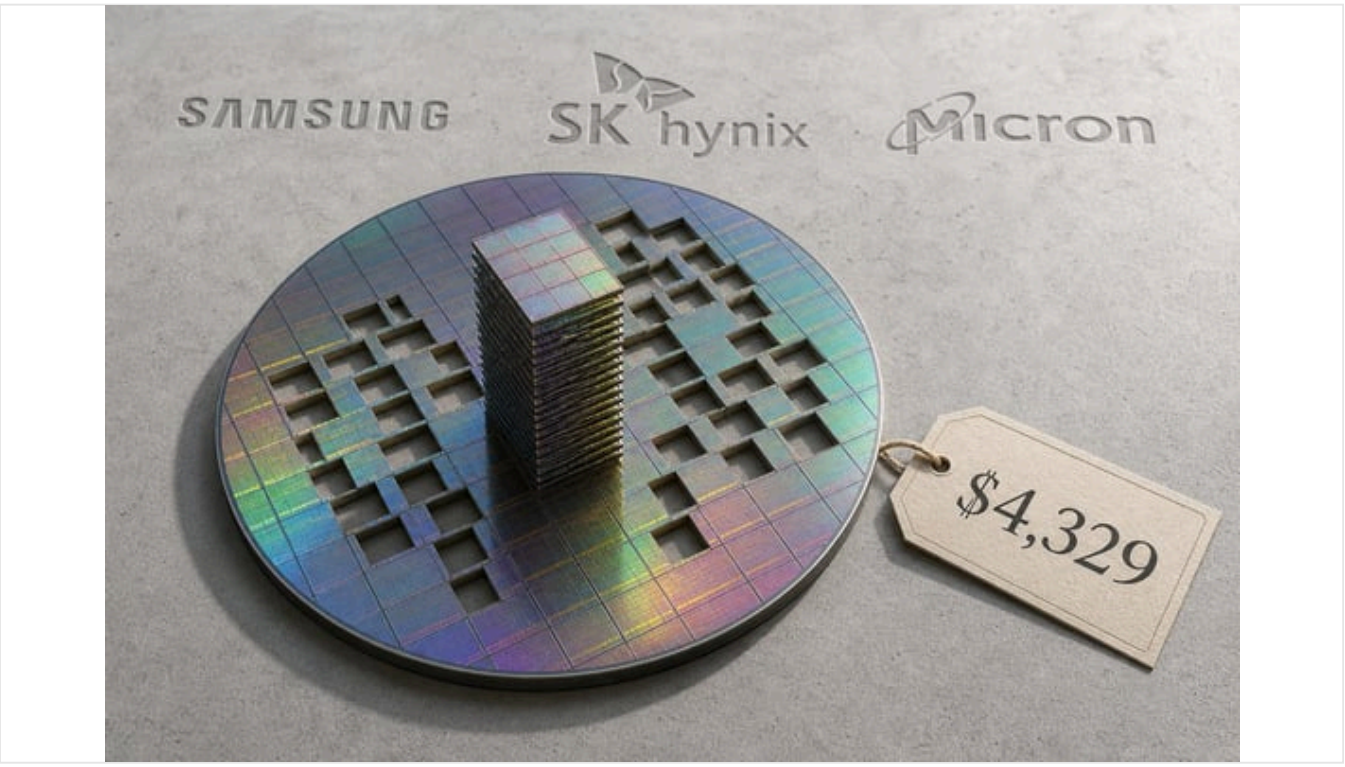
Why AI GPU Prices Are Rising in 2026: HBM Shortage Explained

Published July 16, 2026 36 min read



You've hit your session limit · resets 8:50am (UTC)This matters for consumer and enterprise GPU buyers alike because HBM, GDDR (the graphics memory used in gaming cards), and conventional DDR5 system memory all draw from the same underlying wafer capacity, produced by the same three companies: Samsung Electronics, SK Hynix, and Micron Technology, which together control roughly 90 percent to 95 percent of global DRAM output (Source: www.techtimes.com). When those companies prioritize wafer allocation toward HBM, which commands far higher margins, every other memory product, from a laptop's RAM to a graphics card's VRAM (video RAM) to a games console's system memory, becomes scarcer and more expensive. CNBC reported in January 2026 that "there won't be enough memory to meet worldwide demand because powerful AI chips made by the likes of Nvidia, AMD and Google need so much of it" (Source: www.cnbc.com).

This report explains, in order, what HBM actually is and how it differs from the GDDR memory in gaming cards; why Samsung, SK Hynix, and Micron sold out their 2026 HBM capacity months in advance; how that shortage has driven double- and triple-digit percentage price increases in conventional DRAM and pushed flagship consumer GPU prices toward $5,000; how the effects have spread to game consoles, laptops, and even antitrust litigation; and what buyers, whether individual consumers or enterprise AI infrastructure teams, can realistically expect through 2027 and 2028. Every figure below is sourced to a named research firm, official corporate filing, or wire report, with an "as of" date, because in a market this volatile, precision about timing matters as much as precision about magnitude.

What HBM, GDDR, and DRAM Actually Are

Understanding the 2026 price surge requires distinguishing between three related but distinct memory technologies that are frequently conflated in casual coverage.

DRAM (dynamic random-access memory) is the foundational technology: a type of memory that stores each bit of data in a capacitor that must be periodically refreshed, used in some form in virtually all computer memory, from a smartphone to a supercomputer. **Conventional DRAM** in this report refers to standard modules such as DDR5 (double data rate, fifth generation) used in servers, PCs, and smartphones (as LPDDR, or low-power DDR).

GDDR (graphics double data rate memory) is a variant of DRAM optimized for graphics cards. Nvidia's RTX 50 series uses **GDDR7**, the newest generation, while AMD's Radeon RX 9000 series uses the prior-generation **GDDR6** (Source: www.techtimes.com). GDDR chips are mounted flat on the graphics card's circuit board around the GPU die, connected by a relatively narrow bus (32 bits per chip, with 16 or more chips per card).

HBM (High Bandwidth Memory) takes a fundamentally different physical approach. Rather than mounting memory chips flat on a board, HBM stacks up to 32 DRAM dies vertically and connects them using through-silicon vias (TSVs), microscopic electrical connections drilled through the silicon itself, and microbumps (Source: en.wikipedia.org). SK Hynix produced the first HBM memory chip in 2013, and the first devices to ship with HBM were AMD's Radeon-brand Fiji GPUs in 2015, meaning HBM predates the current AI boom by more than a decade even though its 2026 supply crunch is entirely AI-driven (Source: en.wikipedia.org). This stack is then connected to the GPU or CPU through a silicon interposer, a specialized substrate that permits an extremely wide data bus, up to 1,024 bits per stack, compared with 32 bits for a single GDDR chip (Source: en.wikipedia.org). This width, not raw clock speed, is why HBM delivers dramatically more bandwidth per package: the JEDEC-standardized progression runs from HBM1 in 2013 at 128 gigabytes per second (GB/s) per stack, through HBM3E at up to 1,229 GB/s, to HBM4, standardized in April 2025, which supports up to 64 gigabytes (GB) of capacity per stack and roughly 2,048 GB/s of bandwidth (Source: en.wikipedia.org).

That bandwidth comes at a manufacturing cost. Because HBM requires many more processing steps (TSV drilling, precision die stacking, interposer bonding) than a flat GDDR or DDR5 module, it consumes far more of a memory maker's finite wafer capacity per gigabyte produced. According to industry analysis cited in trade press, producing a single gigabyte of HBM consumes roughly four gigabytes' worth of equivalent standard DRAM wafer area (Source: www.techtimes.com), while Micron itself has described the ratio as a "three-to-one basis" in comments to CNBC (Source: www.cnbc.com). Whichever exact ratio is used, the underlying economics are the same: HBM is several times more wafer-intensive than the memory used in a gaming PC, and because AI data-center demand for it is effectively uncapped, every additional wafer diverted to HBM production is wafer capacity permanently unavailable for GDDR, DDR5, or LPDDR production (Source: en.wikipedia.org).

Why Samsung, SK Hynix, and Micron Sold Out Their 2026 HBM Capacity

The shortage did not emerge gradually; it was announced, in effect, in advance. As early as October 2025, SK Hynix's head of DRAM marketing, Kim Kyu-hyun, told the Financial Times that "our DRAM, NAND, as well as HBM capacity for next year has been sold out" (Source: www.techspot.com). SK Hynix's own third-quarter 2025 results confirmed that "HBM supply discussion for next year completed; HBM4 shipments to begin in Q4 this year, demand for all DRAM and NAND products secured for next year" (Source: news.skhynix.com). That quarter, SK Hynix posted an operating profit of 11.4 trillion won ($8 billion), a 62 percent increase from the same period a year earlier, driven by "strong orders for its high-bandwidth memory chips" (Source: www.techspot.com). Research firm TrendForce notes that SK Hynix alone "controls more than half of the global HBM market," while TechSpot separately reported that "rival Samsung Electronics holds just over a quarter of the market, while Micron Technology in the United States captures much of the remainder" as of late 2025 (Source: www.techspot.com) (Source: www.techspot.com); by the second quarter of 2025, SK Hynix's share stood at 62 percent, Micron's at 21 percent, and Samsung's at 17 percent according to Counterpoint Research figures reported by Chosun Biz (Source: www.astutegroup.com).

Micron followed the same script. The company forecast in September 2025 that HBM revenue would reach an annualized run-rate around $8 billion, and Reuters reported that Micron's "HBM revenue grew to nearly $2 billion" in the prior quarter alone as the company confirmed that "pricing for HBM will increase with advancing version" (Source: www.reuters.com) (Source: www.reuters.com), and later confirmed that its full HBM4 capacity for 2026 was sold out under fixed-price contracts (Source: www.astutegroup.com). By its fiscal third quarter of 2026 (reported June 24, 2026), Micron's revenue had climbed to $41.46 billion, versus $23.86 billion in the prior quarter and $9.30 billion a year earlier (Source: www.globenewswire.com), with GAAP gross margin expanding to 84.6 percent of revenue from 37.7 percent a year earlier (Source: www.globenewswire.com). Micron CEO Sanjay Mehrotra said the results reflect "the strategic value of memory in the AI era," adding that "multi-year Strategic Customer Agreements will significantly enhance the durability and predictability of Micron's strong financial performance" (Source: www.globenewswire.com). Micron's own Q3 FY26 release confirmed that HBM4, "built on 1-beta DRAM technology, is in high-volume shipments for our lead customer's platform" (Source: www.globenewswire.com).

Samsung, meanwhile, reported its highest-ever quarterly consolidated revenue of KRW93.8 trillion in the fourth quarter of 2025, a 9 percent sequential increase, with operating profit also at an all-time high of KRW20.1 trillion (Source: news.samsung.com). Samsung's own release states that its "Memory Business" set an "all-time high for quarterly revenue and operating profit, driven by expanded sales of HBM and other high-value-added products, as well as the overall market price surge" (Source: news.samsung.com). Samsung added that it aims to begin "delivering HBM4 products this quarter, including those with industry-leading 11.7Gbps performance, with the aim of reestablishing a leadership position in the high-end HBM market" (Source: news.samsung.com).

Behind these individual results lies a structural, industry-wide shift documented by IDC: the memory market has reached "an unprecedented inflexion point, with demand materially outpacing supply," driven by a reallocation of manufacturing capacity "away from consumer electronics toward high-margin memory solutions to support AI" (Source: www.idc.com). IDC characterizes this as "not just a cyclical shortage driven by a mismatch in supply and demand, but a potentially permanent, strategic reallocation of the world's silicon wafer capacity" (Source: www.idc.com). Demand for AI chips is itself diversifying beyond GPUs: Goldman Sachs forecast in mid-2025 that HBM demand from GPUs would grow 23 percent year-over-year in 2026, while demand from custom AI ASICs (application-specific integrated circuits, used by hyperscalers such as Google's TPU line) would soar 82 percent, reaching 33 percent of total HBM demand (Source: www.trendforce.com). TrendForce further estimates that HBM's share of total DRAM wafer input among the top three suppliers will rise from approximately 18 percent at the end of 2025 to 22 percent by the end of 2026 and 30 percent by the end of 2027 (Source: www.trendforce.com), a trajectory that leaves progressively less capacity for every other memory product each year.

How the Shortage Reaches DRAM Contract Prices

The HBM capacity squeeze does not stay contained to AI accelerators; it pushes through to the price of every DRAM product because suppliers reallocate wafer capacity toward whichever product line offers the best margin. TrendForce's data illustrates the scale precisely. In early January 2026, the firm forecast conventional DRAM contract prices would rise 55 percent to 60 percent quarter-over-quarter in the first quarter; within a month it had revised that estimate sharply upward to 90 percent to 95 percent (Source: www.trendforce.com). Server DRAM prices specifically were expected to climb "around 90% QoQ in 1Q26, marking the largest quarterly increase on record" (Source: www.trendforce.com), while PC DRAM prices were projected to more than double, "increase by over 100% QoQ, setting a new record for a quarterly surge" (Source: www.trendforce.com).

Actual first-quarter 2026 results, reported by TrendForce in June, came in close to those revised forecasts, with total industry DRAM revenue jumping 81 percent quarter-over-quarter to $97 billion. Samsung's revenue rose 93.4 percent quarter-over-quarter to $37.32 billion, lifting its DRAM market share to 38.5 percent (Source: www.trendforce.com); SK Hynix's revenue rose 62.5 percent to $27.98 billion, holding a 28.8 percent share (Source: www.trendforce.com); and Micron's revenue surged 81.6 percent to $21.75 billion, holding a 22.4 percent share (Source: www.trendforce.com). Wikipedia's technical summary of the period describes DRAM and NAND prices as having "experienced compounded increases, some exceeding 200%, since early 2025" because of "unprecedented demand coming from the AI sector" (Source: en.wikipedia.org).

The effect on desktop consumers has been visible in retail RAM prices. A June 2026 class-action lawsuit against Samsung, SK Hynix, and Micron alleges the companies drove the cost of conventional DRAM up by approximately 700 percent since 2022 by coordinating a "pivot" to HBM as justification for restricting older memory formats (Source: finance.yahoo.com). PCMag reported that consumers were "seeing prices for 64GB DDR5 RAM kits exceeding those of many midrange desktop processors" as of mid-2026 (Source: uk.pcmag.com), a striking role reversal in a market where memory has traditionally been one of the cheapest components in a PC build. On Reddit's r/pcmasterrace community, users have described the trend bluntly: one commenter on a thread discussing Nvidia's production cuts wrote that "these higher prices don't go away, that's just the new price now" (Source: www.reddit.com).

There are early signs of deceleration, though not reversal. TrendForce's outlook for the third quarter of 2026 calls for server DRAM contract prices to rise a comparatively modest 13 percent to 18 percent quarter-over-quarter, as long-term agreements begin to cap the pace of increases (Source: www.trendforce.com). That is still an increase, not a decline, and TrendForce separately forecasts SLC NAND (single-level-cell flash storage) prices climbing 120 percent to 170 percent in the second half of 2026 as manufacturers continue reallocating production lines toward higher-margin server and AI products (Source: www.trendforce.com).

How the Shortage Reaches Consumer Graphics Cards

Consumer GPUs sit at the far end of the same supply chain that feeds AI data centers, and the mechanism connecting the two is direct: Nvidia's RTX 50 series and AMD's Radeon RX 9000 series both depend on GDDR memory manufactured on the same production lines, by the same three companies, that produce HBM. When TrendForce and Kbench reported in January 2026 that Nvidia and AMD were preparing phased price increases across their full GPU lineups, the reporting was explicit that "demand for memory used in AI data centers has far exceeded supply, driving prices of GDDR7 and GDDR6 up by several hundred percent within a matter of months" (Source: www.trendforce.com), with memory now accounting for "more than 80% of the total bill of materials (BOM) for GPUs" (Source: www.trendforce.com). By mid-2026, that same 80 percent-of-bill-of-materials figure was independently corroborated by Tech Times' reporting on high-end cards (Source: www.techtimes.com), which reported that the cost of 16GB of GDDR7 climbed "from roughly $65 to $80 per card in mid-2025 to more than $200 by year-end," driven by spot-market sourcing after long-term supply contracts expired (Source: www.techtimes.com). On a flagship card carrying 32GB of GDDR7, that per-gigabyte inflation scales brutally: Tech Times estimated that VRAM alone accounts for approximately $820 of the RTX 5090's current street price (Source: www.techtimes.com).

The trajectory of the RTX 5090 is the clearest illustration of the pattern. It launched at $1,999 in early 2025; by January 2026, industry sources cited by Kbench suggested it could climb "as high as $5,000 later in 2026" (Source: www.trendforce.com); several add-in-board partners including ASUS and MSI had already pushed custom RTX 5090 cards above $3,000 to $3,500 by that point (Source: www.trendforce.com). By mid-July 2026, the RTX 5090 listed for $4,329 on Amazon, with premium board-partner configurations exceeding $5,000 (Source: www.techtimes.com). Nvidia's own SEC filings corroborate the pressure from the demand side, if not the specific retail price: the company's Edge Computing segment (which now includes gaming alongside workstations, PCs, and robotics) generated $6.4 billion in its fiscal first quarter of 2027, up just 29 percent year-over-year, a fraction of Data Center's 92 percent growth, with the filing noting that consumer PC demand was "tempered by elevated memory and systems prices" (Source: www.sec.gov).

Mid-tier cards have been squeezed as well, though less dramatically. TrendForce reported that RTX 50 models equipped with 16GB or more of VRAM recorded price hikes of roughly 15 percent to 20 percent, while AMD's RX 9000 series saw spot-market premiums of 10 percent to 18 percent in Europe and China (Source: www.trendforce.com) (Source: www.trendforce.com). The **RTX 5060 Ti 16GB**, which launched at $429, was reportedly selling for $489 to $530 or more within weeks of launch (Source: www.techtimes.com). Nvidia has responded by prioritizing lower-memory SKUs, tilting supply toward the 8GB RTX 5060 and RTX 5060 Ti to keep mainstream price points more attainable, while cutting production of higher-VRAM cards by figures reported to range from 20 percent to as much as 40 percent in the first half of 2026 (Source: www.techtimes.com). Even Nvidia's flagship AI accelerator has not been immune to input-cost inflation: pricing for the H200 data-center GPU, which typically sells for $30,000 to $40,000, was expected to rise further in 2026 because each unit carries six stacks of HBM3E whose supply price had increased by about 20 percent (Source: www.trendforce.com).

Perhaps the clearest sign of how severe the constraint has become is what has not happened: as of late June 2026, no desktop RTX 50 "Super" refresh, the kind of mid-cycle, higher-memory variant Nvidia has historically used to reset a generation's value proposition, had been confirmed (Source: www.trendforce.com). The planned refresh would have required higher-density, 3GB-per-chip GDDR7 modules, precisely the component already in shortest supply, illustrating how the shortage has become self-reinforcing: the product that would have eased VRAM scarcity for consumers required the exact input that scarcity had made unavailable.

Beyond Graphics Cards: Consoles, PCs, and the Antitrust Response

The memory shortage's reach extends well past graphics cards into the broader consumer electronics ecosystem, and into the courts. TrendForce data reported by Notebookcheck indicates that memory now accounts for 21 percent to 23 percent of a **Nintendo Switch 2**'s bill of materials and roughly 35 percent for a **PlayStation 5** or Xbox console (Source: www.notebookcheck.net). TrendForce projected that rising memory costs could contribute to a 4.4 percent decline in global console shipments in 2026 compared with the prior year, a figure that had itself been revised up from an earlier 3.5 percent projected decline as the memory crisis intensified (Source: www.notebookcheck.net).

Smartphones face a comparable, if slightly smaller, effect. IDC's downside scenarios show the global smartphone market contracting by 2.9 percent in a moderate case or as much as 5.2 percent in a pessimistic case during 2026, with smartphone average selling prices rising by "3% to 5% in the moderate scenario, or by 6% to 8% in the pessimistic scenario" (Source: www.idc.com). The PC industry faces the same pressure from a different angle. IDC reports that Lenovo, Dell, HP, Acer, and ASUS have all "warned clients of tougher conditions ahead, confirming 15-20% hikes and contract resets as an industry-wide response" heading into the second half of 2026 (Source: www.idc.com). IDC's downside scenarios for the PC market itself range from a 4.9 percent unit-shipment contraction in a moderate case to an 8.9 percent contraction in a pessimistic case, alongside average selling price increases of 4 percent to 8 percent depending on scenario (Source: www.idc.com) (Source: www.idc.com). Apple has begun passing costs through as well: the company stated that "we have never seen a component price increase this much, this quickly," adding, "we have shielded our customers from these increases so far, but we have now reached a point where we need to begin raising prices on a number of products, including today's increases for iPad and Mac" (Source: finance.yahoo.com).

The most consequential legal development came on June 25, 2026, when seventeen named plaintiffs filed a class-action antitrust lawsuit in the U.S. District Court for the Northern District of California against Samsung, SK Hynix, and Micron under the Sherman Antitrust Act (Source: www.techtimes.com). The complaint alleges the three firms "coordinated to restrict DRAM supply and inflate conventional memory prices, citing increases of roughly 700 percent over four years" (Source: www.techtimes.com), and specifically claims the companies used their "coordinated 'pivot' to AI-critical High-Bandwidth Memory (HBM)... as justification to limit the production of older memory formats" (Source: finance.yahoo.com). These are, importantly, unproven allegations, and any legal resolution is likely years away.

The lawsuit invokes a genuinely relevant historical precedent: Samsung and SK Hynix (then Hynix Semiconductor) each pled guilty to criminal DRAM price-fixing charges in the early 2000s, with Samsung paying a $300 million fine and Hynix paying $185 million (Source: www.techtimes.com), while Micron cooperated with Department of Justice investigators as a leniency applicant in that earlier case and avoided criminal fines (Source: www.techtimes.com). Micron has denied the current allegations, stating that it competes "vigorously, fairly and in compliance with all applicable laws

wherever we do business," and that it "will defend ourselves against these claims" (Source: finance.yahoo.com). What is not in dispute is the underlying market structure: three firms controlling roughly 90 percent to 95 percent of global DRAM output constitute a highly concentrated market in which, whether or not any coordination occurred, price increases can propagate faster and further than in a more fragmented industry (Source: www.techtimes.com).

Practical Guidance for Buyers and Infrastructure Planners

Given the structural nature of the shortage, both individual consumers and enterprise buyers face a similar set of tradeoffs. The following considerations apply broadly as of mid-2026:

- **Timing purchases:** Because published quarterly forecasts have consistently been revised upward rather than downward through 2026, as the deceleration only visible by the third quarter shows, deferring a memory-heavy purchase in the hope of a near-term price drop has generally not been rewarded; buyers with an immediate need have tended to fare better acting sooner rather than later.
- **Prioritizing memory capacity over marginal CPU or GPU performance gains**, since VRAM and system RAM, not processor throughput, have driven the bulk of 2026 price inflation (Source: www.techtimes.com).
- **Considering lower-VRAM SKUs** such as the 8GB RTX 5060 or RTX 5060 Ti, which Nvidia has prioritized for supply and which have generally seen smaller price increases than 16GB-plus models, as documented in the pricing patterns above.
- **Evaluating the secondhand and prior-generation GPU market**, which analysts expect to hold value unusually well while new-card pricing remains elevated, since buyers priced out of current-generation cards will bid up demand for older stock.
- **Locking in long-term supply agreements for enterprise buyers**, following the model memory makers themselves have used: Micron's multi-year Strategic Customer Agreements were cited by its own CEO as a source of "durability and predictability" that competitors without such contracts do not enjoy (Source: www.globenewswire.com).
- **Budgeting for elevated total cost of ownership on AI infrastructure**, not just GPU unit price, given that Nvidia's own data-center inventory and supply commitments reached $119.0 billion as of its fiscal first quarter of 2027, roughly a quarter higher than the prior quarter, reflecting the scale of forward purchasing required to secure supply (Source: www.sec.gov).
- **Monitoring quarterly TrendForce and IDC releases directly**, since both organizations publish updated contract-price and shipment forecasts on a roughly monthly to quarterly cadence and have proven more reliable leading indicators than retail-price anecdotes.

Data Analysis and Evidence

The quantitative record of the 2026 memory shortage is unusually well documented because TrendForce, IDC, and the memory makers themselves publish frequent, dated figures. *Table 1* below summarizes the trajectory of conventional DRAM contract price changes across 2026 as tracked by TrendForce, illustrating both the severity of the early-year spike and the tentative deceleration visible by the third quarter.

Table 1: TrendForce Conventional DRAM Contract Price Change by Quarter, 2026

QUARTER	QOQ CONTRACT PRICE CHANGE	NOTES	SOURCE
1Q26 (initial forecast, Jan. 2026)	+55% to 60%	First official 2026 forecast	(Source: www.trendforce.com)
1Q26 (revised forecast, Feb. 2026)	+90% to 95%	Revised sharply upward within one month	(Source: www.trendforce.com)
1Q26 (actual, reported June 2026)	+93% to 98%	Industry DRAM revenue up 81% QoQ to $97B	TrendForce (see body text above)
2Q26 (forecast)	+58% to 63%	Server DRAM allocation continues to squeeze PC/mobile supply	TrendForce (see body text above)
3Q26 (forecast)	+13% to 18%	Long-term agreements begin capping increases	(Source: www.trendforce.com)

Table 1 shows a pattern typical of an acute supply shock rather than a gradual cyclical adjustment: forecasts were revised upward, not downward, during the worst of the crisis, actual outcomes matched or exceeded the revised forecasts, and only by the third quarter did the pace of increase meaningfully slow, without reversing. This matters for interpretation because it indicates suppliers, not just spot-market speculators, were surprised by the strength of AI-driven demand, and that the deceleration visible in the third-quarter figures reflects contractual caps rather than a genuine easing of the underlying supply-demand imbalance.

Table 2 below compares launch and mid-2026 street pricing for the consumer GPUs most affected by the memory shortage, alongside the primary driver cited for each increase.

Table 2: Consumer GPU Pricing, Launch vs. Mid-2026

MODEL	LAUNCH PRICE	MID-2026 STREET PRICE	APPROX. CHANGE	SOURCE
Nvidia RTX 5090	$1,999	$4,329 (some listings above $5,000)	+117% or more	(Source: www.techtimes.com)
Nvidia RTX 5060 Ti 16GB	$429	$489 to $530+	+14% to 24% within weeks	(Source: www.techtimes.com)
RTX 50 series (16GB+ VRAM, general)	Varies by SKU	+15% to 20% board-partner markup	+15% to 20%	TrendForce (see body text above)
AMD RX 9000 series (spot market, Europe/China)	Varies by SKU	+10% to 18% spot premium	+10% to 18%	TrendForce (see body text above)
Nvidia H200 (data-center accelerator)	$30,000 to $40,000 (baseline range)	Further increase expected	HBM3E input cost +~20%	TrendForce (see body text above)

The pattern in *Table 2* underscores a point made throughout this report: price increases scale with memory content, not with silicon complexity. The RTX 5090, which carries the most VRAM of any consumer card at 32GB, saw the largest percentage increase, while the H200, an enterprise-tier product whose price already embeds substantial margin, saw a comparatively modest input-cost increase passed through more gradually because of its longer-term supply contracts.

Underlying all of these figures is the capacity math published by IDC and TrendForce. IDC forecasts 2026 global DRAM supply growth of just 16 percent year-over-year and NAND supply growth of 17 percent year-over-year, both "below historical norms" (Source: www.idc.com), against demand growth driven by AI data centers projected to consume roughly 70 percent of world memory output in 2026 (Source: www.techtimes.com). TrendForce estimates HBM's share of total DRAM bit supply will rise from approximately 8 percent at the end of 2025 to 9 percent by the end of 2026 and 13 percent by the end of 2027 (Source: www.trendforce.com), a modest-looking bit-share figure that understates HBM's true impact because of its much higher wafer intensity per bit produced. Demand growth is compounding the constraint: Nvidia's Rubin Ultra platform, expected in 2027, will require up to 384GB of HBM capacity per GPU, up from 96GB to 192GB in current-generation chips (Source: www.trendforce.com), a roughly two-to-four-fold increase in per-chip memory content that will place further pressure on wafer allocation even if unit shipment growth were to slow.

Case Studies and Real-World Examples

SK Hynix: The First Mover That Sold Out Fastest

SK Hynix's position as the dominant HBM supplier, controlling more than half of the global market as of late 2025 (Source: www.techspot.com), made it the first of the three majors to declare 2026 supply exhausted. Its confirmation that "demand for all DRAM and NAND products secured for next year" (Source: news.skhynix.com) came alongside plans to expand production capacity through a new fabrication extension called **M15X**, where the company opened a cleanroom ahead of schedule and began equipment installation (Source: news.skhynix.com). The company's investment for the following year was expected to rise accordingly, "reflecting the company's continued commitment to a market-aligned investment strategy" (Source: news.skhynix.com). SK Hynix's HBM4 chips began shipping in the fourth quarter of 2025, ahead of full-scale sales expansion planned for 2026 (Source: news.skhynix.com).

Micron: From Commodity Cycles to Contracted Certainty

Micron's transformation illustrates how thoroughly the AI memory boom has changed the economics of what was historically a highly cyclical, commodity-like business. CNBC reported that Micron's "stock is up 247% over the past year, and the company reported that net income nearly tripled" in the quarter immediately preceding the worst of the 2026 shortage (Source: www.cnbc.com). The company's fiscal third-quarter 2026 revenue of $41.46 billion represented more than a fourfold increase from the $9.30 billion reported in the same quarter a year earlier (Source: www.globenewswire.com), while its Cloud Memory Business Unit alone generated $13.77 billion in revenue at an 83 percent gross margin (Source: www.globenewswire.com). Micron guided fourth-quarter fiscal 2026 revenue to $50.0 billion, plus or minus $1.0 billion, with gross margin approaching 86 percent (Source: www.globenewswire.com). Crucially, Micron's own release notes that HBM4E, its next-generation product built on "1-gamma DRAM technology," has volume production expected only in calendar 2027 (Source: www.globenewswire.com), underscoring that meaningful new supply is still roughly a year away even from the most aggressive producer.

Samsung Electronics: Racing to Reclaim HBM Leadership

Samsung's case illustrates that even a company with Samsung's scale and long customer relationships was, for a period, losing ground in the highest-margin segment of the memory market: it held only 17 percent HBM share in the second quarter of 2025, trailing both SK Hynix and Micron (Source: www.astutegroup.com). Its response combined record conventional-DRAM pricing gains with an aggressive HBM4 push: the company resumed construction on its Pyeongtaek P5 fabrication facility (Source: www.astutegroup.com) and targeted delivery of HBM4 products with "industry-leading 11.7Gbps performance" in the first quarter of 2026 (Source: news.samsung.com). By the fourth quarter of 2025, Samsung's Device Solutions division had already posted a 33 percent quarter-over-quarter sales increase (Source: news.samsung.com), evidence that its turnaround strategy was already generating financial results even before its HBM4 ramp reached full scale. Reuters Breakingviews concluded that despite lagging its rivals, Samsung's "financial scale, long-standing customer ties and next-generation roadmap" meant the company "still has a fighting shot" at regaining AI memory share (Source: www.astutegroup.com).

Nvidia: Record Data-Center Growth, Constrained Gaming Segment

Nvidia's own quarterly filings provide a rare, audited view into how the same company can simultaneously be the largest beneficiary of the AI memory boom and one of the parties most publicly affected by its consumer-facing costs. Data Center revenue reached a record $75.2 billion in Nvidia's fiscal first quarter of 2027, up 92 percent year-over-year and 21 percent sequentially (Source: www.sec.gov), while total company revenue hit a record $81.6 billion (Source: www.sec.gov). By contrast, the prior quarter's CFO commentary explicitly flagged that gaming revenue, while up 47 percent year-over-year on strong Blackwell architecture demand, had declined 13 percent sequentially, and that the company expected "supply constraints to be a headwind to Gaming in the first quarter of fiscal 2027 and beyond" (Source: www.sec.gov). Board partner **Gigabyte** summarized Nvidia's evident strategic priority in terms that were not meant as flattery: the company's approach, Gigabyte said, has become "revenue per gigabyte" (Source: www.techtimes.com), a pointed description of how every gigabyte of scarce memory now yields more profit when routed to AI accelerators than when routed to gaming cards.

The DRAM Antitrust Litigation

Whatever its ultimate legal outcome, the June 2026 class-action lawsuit against Samsung, SK Hynix, and Micron functions as a real-world case study in how quickly a supply shock this severe generates formal legal scrutiny. The California complaint calls the three companies "oligopolists" in the DRAM market and alleges "concerted anticompetitive behavior" since 2022 (Source: finance.yahoo.com), seeking a court order to "remedy the ongoing effects of Defendants' unlawful and anticompetitive conduct, including requiring Defendants to cease coordinated supply restriction and restore competitive conditions in the conventional DRAM market" (Source: finance.yahoo.com). Trade associations have separately warned that the price shock could affect manufacturers of vehicles, medical devices, and other goods well beyond computing and gaming (Source: finance.yahoo.com), a reminder that DRAM is a genuinely general-purpose input, not one confined to computing and AI.

Implications and Future Directions

The evidence assembled in this report points to a shortage that is structural rather than transient, and the memory makers' own capital investment plans confirm that view. Micron is building two new fabrication facilities in Boise, Idaho, expected to begin producing memory only in 2027 and 2028, and has additionally broken ground on a new fab in Clay, New York, expected online around 2030 (Source: www.cnbc.com). Because new wafer

fabrication capacity of this kind typically requires years and billions of dollars to bring online, IDC's framing of the shortage as a "potentially permanent, strategic reallocation" rather than a cyclical one (Source: www.idc.com) appears well supported by the capital expenditure timelines the suppliers themselves have published.

Demand is not expected to plateau in the interim. TrendForce projects that HBM capacity per AI chip will rise from a current 96GB to 192GB range to 216GB to 288GB in 2026, driven initially by AI ASIC platforms, before Nvidia's Rubin Ultra architecture pushes per-GPU HBM capacity to 384GB in 2027 (Source: www.trendforce.com). TrendForce expects this to intensify, not ease, the "crowding-out effect on conventional DRAM capacity" and to "strengthen suppliers' pricing power in HBM negotiations" for 2027 (Source: www.trendforce.com). At the same time, there are early signs that HBM pricing power specifically, as distinct from conventional DRAM pricing power, may soften somewhat: TrendForce notes that HBM wafer revenue was actually overtaken by DDR5 64GB RDIMM (registered dual in-line memory module) revenue in the first quarter of 2026, meaning HBM profitability temporarily fell below that of high-capacity server DRAM (Source: www.trendforce.com), a reminder that within the broader shortage, relative profitability across product lines continues to shift and could influence which segments suppliers prioritize next.

For consumer GPU buyers and gaming hardware makers more broadly, the practical implication is that elevated pricing is likely to persist through most of 2027, with the earliest plausible relief tied to new fabrication capacity reaching volume production, which the memory makers' own guidance places no earlier than 2027 and, for some facilities, as late as 2028 to 2030 (Source: www.cnbc.com). For enterprise AI infrastructure planners, the implication is that GPU total cost of ownership should be modeled with continued memory-driven inflation as a baseline assumption rather than a temporary anomaly, particularly for HBM-dependent training and inference hardware where suppliers have signaled they intend to maintain pricing power well into 2027.

Antitrust litigation adds a further layer of uncertainty. Should the June 2026 lawsuit against Samsung, SK Hynix, and Micron proceed to discovery or trial, it could, depending on its outcome, either validate the memory makers' position that current pricing reflects legitimate scarcity and demand, or impose remedies that reshape how the three companies allocate capacity between HBM and conventional DRAM going forward. Given that comparable DRAM price-fixing litigation in the early 2000s took years to resolve and produced fines well below the scale of the damages alleged in the current suit (Source: www.techtimes.com), buyers and infrastructure planners should not expect litigation to alter near-term market dynamics.

Frequently Asked Questions (FAQs)

Why are AI GPU prices rising in 2026? AI GPU prices are rising primarily because the memory chips that surround the GPU die, not the processor itself, have become scarce. Samsung, SK Hynix, and Micron sold out their 2026 HBM production capacity to AI data-center customers months in advance (Source: www.techspot.com), and because HBM and consumer graphics memory draw on the same finite wafer capacity, the shortage has pushed conventional DRAM and GDDR prices sharply higher as well (Source: www.idc.com).

What is causing the HBM memory shortage in 2026? The shortage stems from AI hyperscalers (Microsoft, Google, Meta, and Amazon among them) demanding far more HBM than the three global producers can supply, combined with HBM's far higher wafer intensity relative to conventional memory, roughly three to four times as much wafer capacity per gigabyte (Source: www.cnbc.com) (Source: www.techtimes.com).

Why is GPU memory so expensive right now? Because VRAM, the GDDR memory on a graphics card, now accounts for more than 80 percent of a high-end GPU's total bill of materials, up from a much smaller share historically, as Tech Times documented that "VRAM now accounts for more than 80 percent of the total bill of materials for some high-end graphics cards" following the 2025-2026 GDDR7 price runup (Source: www.techtimes.com).

How much have Nvidia GPU prices increased in 2026? The RTX 5090 rose from a $1,999 launch price to $4,329 on Amazon by mid-July 2026, with some listings exceeding $5,000, while cards with 16GB or more of VRAM broadly saw double-digit percentage price hikes as board partners passed through GDDR7 cost increases (Source: www.techtimes.com).

What are the main causes of the HBM3E supply shortage specifically? HBM3E was the dominant AI accelerator memory standard through 2025 and into 2026, and its shortage traces to the same wafer-allocation dynamics affecting HBM4: Nvidia's H200 uses six stacks of HBM3E per unit, and demand for that generation outstripped available capacity even as HBM4 ramped up alongside it.

How much production capacity do Samsung, SK Hynix, and Micron have for HBM4? All three companies confirmed their entire 2026 HBM4 production capacity was sold out to customers through long-term contracts well before the year began; Micron's HBM4 is described as being in "high-volume shipments for our lead customer's platform" as of mid-2026 (Source: www.globenewswire.com), while SK Hynix's HBM4 shipments began in the fourth quarter of 2025 (Source: news.skhynix.com).

When will GPU prices go down? Most analyst commentary places meaningful relief no earlier than 2027, tied directly to new fabrication capacity reaching volume production; Micron's own new fabs in Boise, Idaho are not expected online until 2027 and 2028 (Source: www.cnbc.com), consistent with IDC's assessment that the reallocation of manufacturing capacity toward AI memory is "potentially permanent" rather than cyclical (Source:

www.idc.com).

What is the difference between HBM and GDDR memory in AI chips? HBM stacks DRAM dies vertically using through-silicon vias to achieve very wide, high-bandwidth connections (up to roughly 2,048 GB/s per HBM4 stack), while GDDR mounts memory chips flat around the GPU with a narrower per-chip bus; HBM is used almost exclusively in data-center AI accelerators, while GDDR is used in consumer and workstation graphics cards (Source: en.wikipedia.org) (Source: www.techtimes.com).

What is the outlook for graphics card prices for the rest of 2026? Prices are expected to remain elevated through the remainder of 2026, with TrendForce forecasting continued, if decelerating, quarterly DRAM contract price increases into the third quarter (Source: www.trendforce.com), and Micron's own capital investment timeline indicating no major new capacity until 2027 and 2028 at the earliest (Source: www.cnbc.com).

Conclusion

The 2026 AI GPU price surge is, at its core, a memory story rather than a processor story. Samsung, SK Hynix, and Micron sold out their entire 2026 HBM production capacity to AI hyperscalers months before the year began, and because HBM, GDDR, and conventional DRAM all compete for the same finite pool of silicon wafer capacity, that scarcity propagated outward into every product these three companies make. Conventional DRAM contract prices rose by roughly 93 percent to 98 percent in a single quarter, consumer flagship GPUs more than doubled in price, game consoles and laptops absorbed double-digit cost increases, and the resulting concentration in a three-firm market has now drawn formal antitrust scrutiny in federal court.

None of this reflects a temporary supply hiccup of the kind the industry has weathered in past cycles. The memory makers' own capital investment disclosures, their multi-year customer contracts, and independent analyst forecasts from TrendForce and IDC all point toward a structural reallocation of global memory production capacity that is likely to persist through 2027 and, in some scenarios, into 2028. Buyers, whether individual consumers weighing a graphics card purchase or enterprise teams budgeting AI infrastructure, should plan around continued elevated pricing rather than an imminent correction. The clearest lesson from the data assembled in this report is that in a market where three companies control the overwhelming majority of a critical input, and where that input's most profitable use case is effectively insatiable in the near term, price signals will keep flowing from the data center to the desktop for the foreseeable future.

Tags: ai gpu prices, hbm memory shortage, gpu prices 2026, nvidia gpu pricing, hbm4, sk hynix, micron hbm, samsung hbm, gddr7 shortage, dram price increase

---