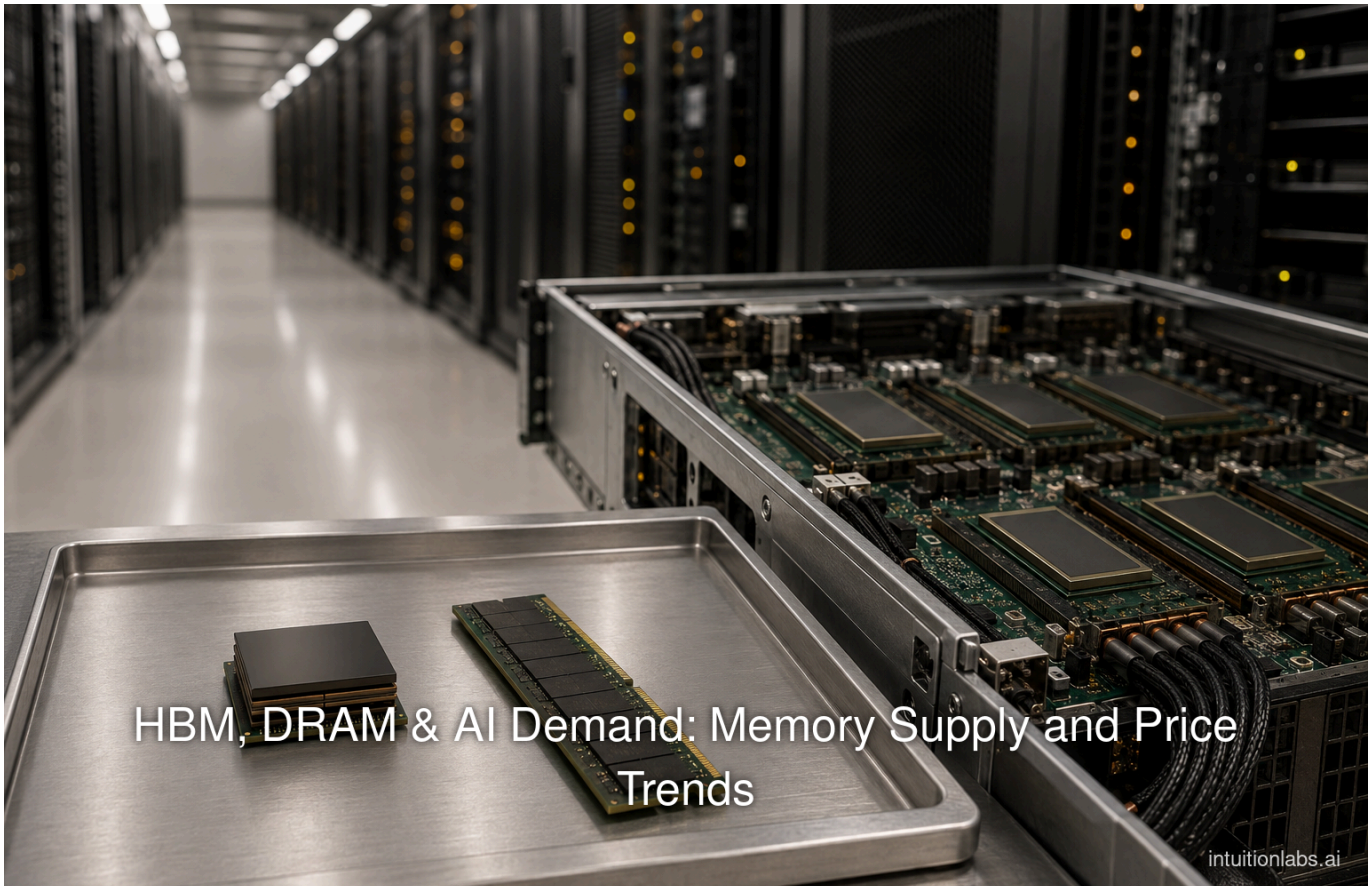
HBM, DRAM & AI Demand: Memory Supply and Price Trends

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · HBM · DRAM · AI memory shortage · HBM3E · HBM4 · memory supply chain · SK hynix · Samsung Electronics · Micron · DRAM pricing



Executive Summary

As of September 2026, the global memory industry is in a documented, multi-quarter supply shortage driven primarily by artificial intelligence (AI) accelerator demand for High Bandwidth Memory (HBM), a specialized memory type that stacks dynamic random-access memory (DRAM) dies vertically to deliver far higher bandwidth than conventional DDR5 memory. TrendForce, the research firm whose quarterly price notes this report treats as the closest available proxy for realized contract pricing, reported that conventional DRAM contract prices rose approximately 93% to 98% quarter over quarter (QoQ) in the first quarter of 2026 alone, driving industry-wide revenue up 81% QoQ to $97 billion (trendforce.com). By September 4, 2026, a 36-gigabyte HBM3E module traded at roughly $2,100 on the spot market, four to five times the $300 to $400 typical long-term-agreement price (fool.com) (247wallst.com).

All three merchant DRAM suppliers, Samsung Electronics, SK hynix, and Micron Technology, describe their 2026 HBM output as effectively committed. Samsung's chief financial officer said on the company's Q1 2026 earnings call that HBM4 sales volume "has already been completely sold out" for the year (en.sedaily.com), and SK hynix said in October 2025 that its "full 2026 offering in both Dram and Nand" had been "spoken for" by customers (straitstimes.com). SK hynix reported second-quarter 2026 operating profit up 557% year over year to a record 60.5 trillion won (news.skhynix.com), while Micron guided fiscal fourth-quarter 2026 revenue to $50.0 billion, up from $9.30 billion a year earlier (stocktitan.net). New capacity, including SK hynix's $4 billion-plus Indiana HBM fab and its 54.3 trillion won ($38.3 billion) Korean fab investment, is not expected to reach volume production before 2028 to 2029 (prnewswire.com) (news.skhynix.com).

Forward-looking forecasts, clearly distinguished throughout this report from realized figures, point to continued tightness rather than resolution: TrendForce's August 2026 note reported analysts expecting HBM contract prices to rise more than 50% in 2027 from 2026 levels, and UBS separately forecast 2027 HBM average selling prices up about 79% (trendforce.com). major PC makers confirmed 15% to 20% price increases (pcworld.com), and retail 32GB DDR5-6000 kits reached $402 in August 2026, up from $110 to $140 a year earlier (tech-insider.org).

This report documents the reproducible method behind each figure, cites JEDEC's own standards timeline for HBM2 through HBM4, and separates AI-hyperscaler, enterprise, and consumer segments to show how a memory type that consumes roughly three times the wafer capacity of DDR5 per gigabyte has reallocated scarce fabrication capacity industry-wide (tomshardware.com). Every 2027 figure surveyed, including from the suppliers themselves, remains an explicit forecast, not a certainty.

Introduction and Background

The global memory industry entered the second half of 2026 in the midst of what suppliers themselves describe as a structural, multi-year supply crunch. Samsung Electronics' global marketing head Wonjin Lee told Bloomberg in January 2026 that the shortage "is going to affect everyone, not just Samsung" (networkworld.com), a statement that anticipated a year of record contract-price increases, sold-out High Bandwidth Memory (HBM) allocations, and knock-on price pressure across personal computers, graphics cards, and enterprise servers. IntuitionLabs previously examined the emerging dynamics of this shortage in a December 2025 report, which described "exponentially rising demand" colliding with manufacturing constraints (intuitionlabs.ai). This report extends that analysis with dated data gathered as of September 2026: reproducible price and capacity figures from the three dynamic random-access memory (DRAM) suppliers, from JEDEC, and from named research firms such as TrendForce, Counterpoint Research, and IDC.

The proximate cause is well documented: artificial intelligence (AI) accelerators, the graphics processing units (GPUs) and custom chips that train and run large AI models, require dramatically more memory bandwidth and capacity than prior generations of computing hardware, and much of that memory takes the form of HBM, a specialized memory type manufactured by stacking DRAM dies vertically and connecting them with through-silicon vias (TSVs). TrendForce's HBM research center estimates that 2025 HBM bit demand grew more than 130% year over year (YoY), with 2026 growth still projected above 70% YoY (trendforce.com). Because HBM production consumes roughly three times the silicon wafer capacity of standard DDR5 (Double Data Rate 5) memory per gigabyte produced, per a Tom's Hardware technical analysis (tomshardware.com), every wafer diverted to HBM production is, in IDC's framing, a wafer denied to smartphone or personal computer (PC) memory (idc.com).

This article separates three categories of claim that are frequently blurred in shortage coverage: dated, realized price and revenue figures reported by suppliers and research firms; forward-looking analyst forecasts, which are explicitly labeled as such; and vendor marketing claims about unreleased products, which are distinguished from independently verified, generally available shipments. It also documents the reproducible method behind each figure so that the underlying calculation, whether a quarter-over-quarter (QoQ) contract price change or a wafer-allocation percentage, can be checked against its original source.

Methodology and Scope

This report synthesizes dated disclosures from three source categories, each weighted differently depending on the claim. **Tier 1 sources** comprise the official investor-relations and press materials of the three merchant DRAM suppliers, Samsung Electronics, SK hynix, and Micron Technology (Nasdaq: MU), plus JEDEC's own standards-publication announcements; these are used for the companies' own reported revenue, margin, and shipment-status figures, and for the underlying technical specifications of each memory standard. **Tier 2 sources** are named research firms, principally TrendForce, whose HBM- and DRAM-focused press center publishes quarterly contract-price trend estimates that this report cites by publication date; where a firm (Counterpoint, IDC, UBS, Bernstein) is quoted only through a secondary outlet because the originator's page could not be fetched, that outlet is named. **Tier 3 and 4 sources** are business wire and trade press, used for consumer- and enterprise-facing price effects that the primary suppliers do not themselves disclose.

Every price or capacity figure below carries the "as of" date its source states, since contract prices, spot prices, and market-share estimates change quarter to quarter. A figure that is an analyst forecast rather than a reported, realized result is labeled explicitly as a forecast. Percentages describing "capacity" or "market share" are as disclosed by the named source; this report does not independently recompute them, since the underlying production data is proprietary. No figure below has been extrapolated by the authors; every number traces to a specific fetched page cited inline.

HBM and DRAM Technical Foundations and the Demand Drivers Behind the Shortage

HBM and conventional DDR-family DRAM are both built from the same basic silicon memory cell, but they are packaged and standardized differently by JEDEC, the industry body that defines memory specifications. JEDEC published the original HBM (JESD235) update in January 2016, defining a device that used "Wide I/O and TSV technologies to support up to 8 GB per device at speeds up to 256 GB/s" (jedec.org), with that 256 GB/s of bandwidth "delivered across a 1024-bit wide device interface that is divided into 8 independent channels" (jedec.org). JEDEC's HBM3 standard (JESD238), published in January 2022, doubled the per-pin data rate of the prior generation, "defining data rates of up to 6.4 Gb/s, equivalent to 819 GB/s per device" (jedec.org) and expanding to 16 independent channels while "supporting 4-high, 8-high and 12-high TSV stacks with provision for a future extension to a 16-high TSV stack" (jedec.org). Nvidia's own technical marketing director, who chaired the JEDEC HBM subcommittee, said at the time that "HBM3 will enable new applications requiring tremendous memory bandwidth and capacity" (jedec.org), foreshadowing the AI-driven demand now straining supply.

HBM3E is codified as JEDEC document JESD238B, its .01 revision "published: Apr 2025" (jedec.org). SK hynix began volume production of the world's first 12-layer HBM3E in September 2024, raising capacity while increasing "the speed of memory operations to 9.6 Gbps, the highest memory speed available today" (news.skhynix.com). Samsung's competing 12-layer HBM3E stack "offers up to 1,180GB/s bandwidth at 9.2Gbps" per pin, according to Samsung's own semiconductor product page (semiconductor.samsung.com). The next generation, **HBM4**, was published by JEDEC as JESD270-4 on April 16, 2025, roughly doubling the interface again: "with transfer speeds up to 8 Gb/s across a 2048-bit interface, HBM4 boosts total bandwidth up to 2 TB/s" (jedec.org) across 32 independent channels, and "supports 4-high, 8-high, 12-high and 16-high DRAM stack configurations with 24 Gb or 32 Gb die densities" (jedec.org). A subsequent JESD270-4A revision, published December 2025, clarifies that "each channel interface maintains a 64 bit data bus operating at double data rate" (jedec.org). By comparison, JEDEC's mainstream DDR5 standard (JESD79-5), published July 14, 2020, was "expected to be launched at 4.8 Gbps" (jedec.org), a fraction of HBM4's per-pin and aggregate bandwidth, since "DDR5 supports double the bandwidth as compared to its predecessor, DDR4" (jedec.org), not an order of magnitude more, as HBM does (see Table 3).

The demand side of the shortage traces directly to AI accelerator design. Nvidia's H100 GPU, the workhorse of the 2023-2024 AI buildout, ships with "GPU Memory 80GB" and bandwidth of "3.35TB/s" per chip according to Nvidia's own product page (nvidia.com), while the dual-GPU H100 NVL variant carries 188GB of HBM3 (nvidia.com). The subsequent Blackwell-generation HGX B200 platform carries "up to 1.4 terabytes (TB) of HBM3E memory" across its eight GPUs (primeline-solutions.com), and a full GB200 NVL72 rack "supports up to 13.5TB of HBM3E" pooled across 72 GPUs (primeline-solutions.com). TrendForce's HBM research finds that per-chip HBM capacity is rising further still, from "96GB/192GB" in earlier AI accelerator generations toward higher configurations as designers race to keep pace with model size (trendforce.com). Epoch AI, analyzing 2025 chip-supply-chain data, found that the four largest AI chip designers "collectively consumed over 90% of global CoWoS packaging capacity and HBM supply by value in 2025" (epoch.ai), versus a far smaller share of advanced logic die production, indicating packaging and HBM, not leading-edge fabrication, were the binding 2025 constraint. OpenAI's chief operating officer Brad Lightcap made the same point directly in public remarks reported in March 2026: "Right now [the bottleneck] is memory. . . . It's been power" (cnas.org). SemiAnalysis, a semiconductor research newsletter, explains the underlying technical reason HBM rather than conventional DRAM serves this role: HBM is "far superior in terms of bandwidth per package than any other form of memory" (newsletter.semianalysis.com), since large AI models are often limited by data movement, not raw compute.

Manufacturer Capacity, Qualification, and Sold-Out Allocations

All three merchant DRAM suppliers, Samsung, SK hynix, and Micron, have publicly described their 2026 HBM output as effectively committed. SK hynix said in October 2025 that its customers had secured its "full 2026 offering in both Dram and Nand spoken for" (straitstimes.com), SK hynix separately announced on September 12, 2025 that "it has completed development and finished preparation of HBM4" for what it called the world's first HBM4 mass-production system (news.skhynix.com), with a per-pin speed that "far exceeded the JEDEC standard operating speed" of 8 Gb/s (news.skhynix.com). Counterpoint Research, cited on SK hynix's own site, credited the company with "62% share of HBM shipments as of Q2 2025 and 57% of revenue as of Q3" (news.skhynix.com), and UBS forecast a 70% SK hynix share of Rubin-platform HBM4 in 2026 (see Table 2). SK hynix's own January 2026 outlook said the market still "expect[s] HBM3E to account for approximately two-thirds of total HBM shipments in 2026" (news.skhynix.com), meaning HBM4 had not yet displaced HBM3E in shipment volume.

To meet this demand, SK hynix's board approved a combined 54 trillion won investment in August 2026 to "invest 35.2 trillion won in Yongin "Y2" fab" for next-generation DRAM and HBM, alongside a Cheongju NAND fab (news.skhynix.com), with the Yongin site "targeting the opening of its first cleanroom in June 2029 to produce high-bandwidth memory" (news.skhynix.com), SK hynix followed with a groundbreaking ceremony on August 27, 2026 for its first HBM production base outside Korea, a West Lafayette, Indiana facility with an "investment of over $4 billion" that "is scheduled to open its cleanroom by October 2028" (prnewswire.com), with mass production targeted for 2029.

Samsung has moved on a parallel track. Nvidia chief executive Jensen Huang confirmed at a June 1, 2026 keynote that "high-bandwidth memory from Samsung Electronics, SK hynix and Micron will go into" Nvidia's next-generation Vera Rubin platform (koreaherald.com), with industry estimates reported by the same outlet putting SK hynix's "share of HBM4 volume for the Rubin launch at roughly 60 to 70 percent," with Samsung taking roughly 25 to 30% and Micron a smaller, supplementary share (koreaherald.com). Samsung said it "cleared Nvidia's qualification tests at the 10 and 11 gigabit-per-second data rates the Rubin design demands" and began shipping 12-layer HBM4 in early 2026 (koreaherald.com). On its Q1 2026 earnings call, Samsung's chief financial officer Park Soon-cheol said the company's HBM4 "sales volume has already been completely sold out" for the year, adding that HBM4 revenue was expected to "exceed half of our total HBM revenue starting in the third quarter" of 2026 (en.sedaily.com) (en.sedaily.com). Samsung displayed HBM4E as its next-generation product at Nvidia's March 2026 GTC conference (detailed in the case study below, an announced capability rather than a currently shipping one). TrendForce reported that Samsung began official HBM4 shipments in February 2026 and was "also targeting an HBM market share of around 38% by year-end" (trendforce.com). Separately, SamMobile reported, citing Samsung's own earnings call, that "Samsung has already sold out of its 2026 supply of HBM4 chips" and that rival suppliers had likewise sold through much of their 2026 supply (sammobile.com).

Micron, the smallest of the three suppliers by DRAM revenue share, said in a June 2025 disclosure that it was already "shipping high-volume HBM to four customers across both GPU and ASIC platforms" and targeted a 24% HBM market share by the end of that year; at that point, TrendForce ranked Micron third in overall DRAM revenue, "capturing 24.3% of the market" behind SK hynix and Samsung (trendforce.com). In its fiscal third-quarter 2026 results (quarter ended May 28, 2026), Micron disclosed that "HBM4, built on 1-beta DRAM technology, is in high-volume shipments for our lead customer's platform" (investors.micron.com), explicitly distinguishing that generally available product from its next-generation HBM4E, whose "volume production" the company said is "expected in calendar 2027" (investors.micron.com). Micron chief executive Sanjay Mehrotra said the company "is investing at record levels in technology, products and supply to address our customers' rapidly growing demand" (investors.micron.com). *Table 2* below summarizes each supplier's disclosed HBM status, distinguishing announced capabilities from shipping, generally available product.

Supplier	HBM generation status (as of stated date)	Disclosed share or allocation	Capacity investment
SK hynix	HBM4 mass-production system completed Sep. 2025 (news.skhynix.com); HBM3E remains ~2/3 of 2026 shipment mix per brokerage consensus (news.skhynix.com)	62% of HBM shipments, Q2 2025, per Counterpoint (news.skhynix.com); UBS forecasts ~70% of Rubin-platform HBM4 in 2026 (forecast, not realized) (news.skhynix.com)	54.3 trillion won (~$38.3B) Yongin/Cheongju fabs, cleanrooms 2028 to 2029 (news.skhynix.com); $4B+ Indiana HBM fab, groundbreaking Aug. 2026 (prnewswire.com)
Samsung	HBM4 qualified at 10 to 11 Gbps for Vera Rubin, shipping since early 2026	HBM4 volume "completely sold out" for 2026 per CFO, Q1 2026	~70% of overall memory production locked through 2031

Supplier	HBM generation status (as of stated date)	Disclosed share or allocation	Capacity investment
	(koreaherald.com); HBM4 in mass production and designed for Vera Rubin; HBM4E displayed Mar. 2026 and samples began shipping in May (Samsung Semiconductor; Samsung Newsroom)	call (en.sedaily.com); targeting ~38% overall HBM share by year-end 2026 (TrendForce) (trendforce.com)	under long-term hyperscaler agreements (247wallst.com)
Micron	HBM4 (1-beta node) in high-volume shipment to lead customer as of fiscal Q3 2026 (investors.micron.com); HBM4E volume production targeted calendar 2027 (announced, not yet shipping) (investors.micron.com)	Targeted 23 to 24% HBM share by end of 2025 (trendforce.com); 24.3% overall DRAM revenue share, Q1 2025 (trendforce.com)	Cloud Memory Business Unit revenue $13.77B in fiscal Q3 2026, up from $3.39B a year earlier (stocktitan.net)

Table 2 makes clear that all three suppliers describe 2026 HBM output in terms of allocation and qualification rather than open-market availability: SK hynix and Samsung each report volumes effectively committed or sold out, Micron describes high-volume shipment concentrated with a single lead customer, and all three suppliers' next-generation (HBM4E) products remain in an announced, pre-volume-production state as of the dates shown. The practical implication for a buyer evaluating AI infrastructure in late 2026 is that available HBM supply is set by long-term agreements negotiated months or years in advance, not by current spot demand.

Pricing Trends: Contract, Spot, and Forecast Data

Dated pricing data shows a market that has repriced sharply upward through 2026, with realized quarterly figures running well ahead of even the suppliers' own earlier guidance. TrendForce, whose quarterly DRAM and NAND price trend notes are treated by the industry as the closest available proxy for actual contract pricing, reported that first-quarter 2026 "conventional DRAM, which increased by approximately 93% to 98% QoQ," drove "overall industry revenue increase of 81% QoQ, reaching $97 billion" for the quarter, with that revenue split across "driving quarterly revenue up 93.4% QoQ to $37.32 billion" for market-leading Samsung, alongside SK hynix and Micron gains of similar order (trendforce.com). Legacy, trailing-edge parts have moved even more sharply: TrendForce estimated legacy DDR2 contract prices would rise approximately 55 to 60% in the second quarter of 2026, followed by a further 35 to 40% increase in the third quarter (see Table 1), reflecting capacity being pulled away from older process nodes toward AI-related products.

By the third quarter of 2026, TrendForce's own price-trend notes described some moderation at the margin, forecasting conventional DRAM contract prices would rise a further "13-18% QoQ" and NAND flash "10-15% QoQ," a deceleration from the prior quarters' near-doubling but still a substantial increase in absolute terms (trendforce.com); server-grade DRAM specifically was forecast to rise "13-18% quarter-over-quarter in 3Q26," even as TrendForce separately forecast that "total RDIMM bit supply will grow by only 15-20% YoY" in 2027, lagging server-CPU shipment growth and implying continued tightness for registered server memory modules (trendforce.com). HBM-specific contract pricing moved in the same direction: Samsung and SK hynix "have raised HBM3E supply prices by nearly 20% for 2026" according to a December 2025 TrendForce report (trendforce.com), a figure TrendForce called unusual for a category typically governed by fixed agreements.

Spot-market pricing, where memory trades for immediate delivery outside long-term agreements, has diverged sharply from those same contract prices. As of September 4, 2026, a 36-gigabyte (GB) HBM3E chip "sells for about $2,100 on the spot market, where chips trade for immediate delivery" (fool.com), roughly four to five times the $300 to $400 typical long-term-agreement price (247wallst.com). Korean customs export data shows the same dynamic in aggregate trade figures: the average HBM export unit price "jumped about 37% in two months, from $16.76 to $22.90" between May and July 2026 even as export volumes fell, according to The Motley Fool's analysis of the underlying trade data (fool.com). Those price dynamics are visible in supplier financial results: SK hynix's DRAM and NAND average selling prices (ASPs) "increased around 30% and 50%, respectively, quarter over quarter in Q2" 2026, per Korea Investment & Securities estimates cited by TrendForce (trendforce.com), a dynamic reflected in SK hynix's own record quarterly results (discussed below).

Looking forward, TrendForce's August 2026 note projected that server DRAM contract prices, having "risen by a cumulative 64% in 2H25, with a further jump of approximately 270% expected in 2026" (a forecast), and analysts expect HBM contract prices to rise "by more than 50% next year from 2026 levels," also a forecast rather than a realized figure; the same note reported that Nvidia is "planning to raise prices for Vera Rubin and Grace Blackwell servers by more than 15% in early 2027" and that "some DDR5 prices have nearly quadrupled over the past year," underscoring that DDR5, not only HBM, has repriced sharply (trendforce.com). UBS, cited in the same TrendForce report, separately forecasts that "2027 HBM average selling prices (ASPs) to rise by about 79% from this year," while Bernstein forecasts HBM4 pricing rising to US$53 per GB by 2027 (see Table 1), and TrendForce separately estimates that memory "could account for around 29% of the Vera Rubin VR200 system's estimated US$2.1 million bill of materials" without design changes, above Nvidia's preferred 20% level (trendforce.com).

Table 1 below consolidates these dated price observations chronologically, distinguishing realized, reported figures from forward-looking forecasts.

Period / Date reported	Metric	Change (realized or forecast)	Source
1Q 2026 (reported Jun. 2026)	Conventional DRAM contract price	+93% to 98% QoQ (realized)	TrendForce (trendforce.com)
2Q 2026 forecast (issued Jun. 2026)	Legacy DDR2 contract price	+55 to 60% QoQ (forecast)	TrendForce
Q2 2026 (reported Jul. 2026)	SK hynix DRAM / NAND ASP	+30% / +50% QoQ (realized)	TrendForce citing Korea Investment & Securities (trendforce.com)
3Q 2026 forecast (issued Jul. 2026)	Conventional DRAM / server DRAM / NAND contract price	+13 to 18% / +13 to 18% / +10 to 15% QoQ (forecast)	TrendForce (trendforce.com) (trendforce.com)
As of Sep. 4, 2026	HBM3E 36GB module, spot vs. long-term agreement	$2,100 spot vs. $300 to $400 LTA (realized)	The Motley Fool / 247wallst.com (fool.com) (247wallst.com)
May to Jul. 2026	Korea HBM export unit price	$16.76 to $22.90, +37% (realized)	The Motley Fool (fool.com)
Aug. 2026 retail (vs. Q3 2025)	32GB DDR5-6000 kit, US retail	$402 vs. $110 to $140 (realized)	Tech-Insider aggregation of Tom's Hardware tracking (tech-insider.org)

Period / Date reported	Metric	Change (realized or forecast)	Source
2026 to 2027 forecast (issued Aug. 2026)	Server DRAM contract price	+64% cumulative 2H25, +~270% in 2026 (forecast)	TrendForce (trendforce.com)
2027 forecast (issued Aug. 2026)	HBM contract price / HBM ASP	+more than 50% / +79% (forecast, two estimates)	TrendForce / UBS (cited above)
2027 forecast (issued Jul. 2026)	HBM4 price per GB	$53/GB (forecast)	Bernstein via TrendForce (trendforce.com)

The chronology in *Table 1* shows realized quarterly increases running in the double and triple digits through most of 2026, moderating somewhat by Q3 on a percentage basis while still compounding on an already elevated base, and it shows every forward 2027 figure explicitly labeled by its own source as a forecast rather than a settled outcome. The spread between spot and long-term-agreement pricing, a factor of roughly four to five for HBM3E as of September 2026, is itself informative: it indicates that suppliers have chosen to allocate the large majority of capacity to multi-year contracts rather than the open market, which caps how much of the shortage's price signal reaches any single buyer without such an agreement.

Data Analysis and Evidence

Aggregate market-size data corroborates the price signals described above. TrendForce, in a May 2026 revision, raised its full-year 2026 global memory market forecast, "increasing its 2026 estimate from US$551.6 billion in the previous report to $889.3 billion," and separately revised its 2027 forecast "upward from $842.7 billion to more than $1.28 trillion, representing annual growth of approximately 44%" (trendforce.com), attributing the raised outlook substantially to AI inference workloads. That trajectory is reflected in supplier-level financial disclosures. SK hynix reported second-quarter 2026 (period ended June 30, 2026) "revenues of 79.3187 trillion won, operating profit of 60.5426 trillion won, net profit of 93.9226 trillion won" (news.skhynix.com), with "revenue and operating profit increased by 257% and 557% year-over-year" respectively (news.skhynix.com). Micron's fiscal third-quarter 2026 results (period ended May 28, 2026) showed "revenue of $41.46 billion versus $23.86 billion for the prior quarter and $9.30 billion for the same period last year" (stocktitan.net), with the company guiding the following quarter to an "FQ4 2026 revenue guidance of $50.0 billion" plus or minus $1.0 billion (stocktitan.net). Samsung Electronics reported consolidated second-quarter 2026 revenue of "KRW 171.5 trillion in consolidated revenue, another all-time quarterly high," up 28% QoQ (news.samsung.com), and said it expects the memory market to stay undersupplied for the remainder of 2026: "this is projected to keep the market undersupplied, despite partial demand moderation in mobile and PCs" (news.samsung.com).

Independent estimates of demand composition help explain why AI-linked memory has outpaced other segments. TrendForce's own HBM-specific modeling estimated that HBM would consume "approximately 18%, 22%, and 30% of total" DRAM wafer input for the top three suppliers by the end of 2025, 2026, and 2027 respectively, a rising share that the firm said reflects a "crowding-out effect on conventional DRAM capacity" that "is expected to intensify" through the period (trendforce.com). As of August 2026, TrendForce forecast that global "HBM bit shipments are projected to grow by 50-60% YoY" in 2027, while also cautioning that "DRAM supply will remain tight through 2027, and uncertainty persists" over whether that shipment growth will be sufficient to close the gap (trendforce.com), meaning that projected growth rate is itself the constrained outcome, not evidence of resolution.

Analysis of Key Segments: Who Is Affected and How

The shortage does not affect every buyer of memory equally, and the evidence supports distinguishing at least three segments. The first is **AI hyperscaler and accelerator demand**, where the largest cloud providers and chipmakers have secured supply years in advance through long-term agreements; Epoch AI's analysis found the four largest AI chip designers "collectively consumed over 90% of global CoWoS packaging capacity and HBM supply by value in 2025" (epoch.ai), meaning most HBM output already flows to a small number of buyers under negotiated terms rather than being available on the open market. This segment experiences the shortage chiefly as a qualification and roadmap risk rather than outright unavailability, since agreements are typically struck before a chip's public launch.

The second segment is **enterprise and general-purpose server buyers** outside the largest AI programs, competing for registered DIMM (RDIMM) and conventional server DRAM without the scale to negotiate long-term agreements. A coalition representing the telecommunications, automotive, medical-device, and retail industries warned US regulators in mid-2026 that "expanding artificial intelligence (AI) data centers consume an enormous share of available memory chip capacity" (datacenterknowledge.com), citing TrendForce data that conventional DRAM contract prices had "rose roughly 93 to 98% quarter over quarter in the first quarter" of 2026 (datacenterknowledge.com) as evidence non-AI industrial buyers were being squeezed. Morgan Stanley research separately found that "memory chip prices have spiked by six times in the past year" as of mid-2026 (devdiscourse.com).

The third segment is **consumer and PC-adjacent demand**, where the shortage has arrived as a retail price shock without the buffer of a long-term agreement. AMD reportedly told supply-chain partners "it will raise graphics cards prices by at least 10% in the new-year due to rising memory prices" (archive.org / tomshardware.com), and PC makers "Lenovo, Dell, HP, Acer and Asus have warned clients of tougher conditions ahead, confirming 15-20 percent hikes" tied to DRAM and solid-state drive shortages (pcworld.com). Retail-market tracking aggregated from Tom's Hardware pricing data shows a mainstream "32GB DDR5-6000 kit at $402" as of August 2026, "up from just $110 to $140 as recently as Q3 2025" (tech-insider.org), and the same aggregation cites a TrendForce forecast revision describing "PC DRAM contract prices to a staggering 105-110% quarter-over-quarter increase," which it called the "steepest single-quarter surge on record" (tech-insider.org).

Samsung itself has acknowledged this cross-segment spillover directly: global marketing head Wonjin Lee told Bloomberg in January 2026 that the coming price surge "is going to affect everyone, not just Samsung" (networkworld.com). The evidence supports a causal reading in which AI-driven HBM and server DRAM demand set the initial capacity reallocation in motion, with consumer PC and component pricing responding with a lag; this report treats the precise size of that lag as unquantified by any source fetched, since no supplier or research firm publishes a formal cross-segment attribution model.

Case Studies and Real-World Examples

SK hynix's Indiana HBM production base. On August 27, 2026, SK hynix held a groundbreaking ceremony for its first HBM production facility outside Korea, in West Lafayette, Indiana, describing an "investment of over $4 billion" with a cleanroom "scheduled to open . . . by October 2028" and mass production of next-generation HBM targeted for the second half of 2029 (prnewswire.com). The facility illustrates the multi-year lead time inherent to HBM

capacity expansion: even a well-capitalized supplier moving with urgency requires roughly three years between groundbreaking and volume output, a timeline that constrains how quickly new capacity investment announced in 2025 and 2026 can translate into available supply.

Samsung's HBM4E unveiling at Nvidia GTC 2026. In March 2026, Samsung used Nvidia's GTC developer conference to present its sixth-generation HBM4 as already "in mass production and . . . designed for the NVIDIA Vera Rubin platform" (semiconductor.samsung.com), alongside a hybrid copper bonding packaging technique intended to enable future 16-plus-layer stacks with over 20% lower heat resistance than thermal compression bonding (semiconductor.samsung.com). This launch illustrates the distinction this report draws throughout between an announced capability, shown at a trade conference, and independently confirmed, broad general availability, which for HBM4E remained pending as of the dates covered here.

Implications and Future Directions

The evidence assembled above supports several forward-looking observations, each qualified by the forecast-versus-realized distinction maintained throughout this report. First, capacity relief is unlikely before 2028 to 2029 at the earliest: both SK hynix's Yongin fab (cleanroom target June 2029) and its Indiana facility (cleanroom target October 2028, volume HBM output second half of 2029) sit multiple years out from this report's September 2026 publication date, meaning the supply response is structurally delayed. Second, pricing power appears likely to persist through 2027 on the evidence available: TrendForce's own August 2026 note forecasts HBM bit shipment growth of 50 to 60% YoY in 2027, which the firm itself frames as insufficient to fully resolve tightness.

For organizations planning private or hybrid AI infrastructure deployments, memory-driven cost inflation is one input among several that now needs to be modeled explicitly rather than assumed away. IntuitionLabs' own analysis of the private-AI data center market found that Gartner estimated "worldwide AI spending totaled $1.765 trillion in 2025" and forecast a rise to $2.596 trillion for 2026 (intuitionlabs.ai), a buildout scale against which memory now represents a rising, not fixed, share of system cost. As an adjacent AI and life-sciences advisory rather than a hardware vendor, IntuitionLabs' role is to help regulated life-sciences organizations account for this memory-driven cost volatility, and the multi-year allocation cycles documented above, in their infrastructure roadmaps and vendor negotiations, rather than to supply hardware itself. Finally, organizations budgeting for on-premises AI hardware in 2027 should treat HBM and DDR5 pricing as independently volatile line items, since the evidence shows the two have moved on related but not identical trajectories.

Conclusion

Dated evidence gathered as of September 2026 shows a memory market reshaped by AI accelerator demand at every level, from JEDEC's own standards roadmap, which has pushed HBM bandwidth from 256 GB/s in 2016 to a targeted 2 TB/s per stack under HBM4, to supplier earnings calls describing 2026 HBM output as sold out, to consumer DDR5 retail prices that nearly quadrupled within a year. Samsung, SK hynix, and Micron have each disclosed record or near-record revenue tied to the shortage, alongside multi-billion-dollar capacity investments whose earliest output is not expected before 2028 to 2029. Realized contract-price increases through 2026 ran from double digits to, for conventional DRAM in the first quarter, nearly 100% quarter over quarter, while every 2027 figure cited here remains an explicitly labeled forecast, not a settled outcome. The clearest throughline in the evidence is that AI accelerator demand for HBM, which consumes roughly three times the wafer capacity of

equivalent DDR5 output, has reallocated scarce fabrication and packaging capacity away from conventional memory, a dynamic that industry and research-firm sources agree is structural rather than a short-term hiccup, even as the pace and endpoint of resolution remain genuinely contested among the forecasts surveyed here.

Frequently Asked Questions (FAQs)

What is causing the HBM and DRAM shortage? The shortage traces primarily to AI accelerators requiring far more memory bandwidth and capacity than prior computing hardware, delivered through HBM, which TrendForce estimates will consume approximately 18%, 22%, and 30% of total DRAM wafer input for the top three suppliers by the end of 2025, 2026, and 2027 respectively (cited above), crowding out capacity previously available for conventional DRAM.

When will DRAM and HBM prices stabilize? No source fetched for this report gives a specific stabilization date. TrendForce's own August 2026 forecast projects continued tightness through 2027, with HBM contract prices potentially rising more than 50% in 2027 (cited above), while SK hynix's own investment timeline suggests meaningful new capacity is not expected before 2028 to 2029 (news.skhynix.com).

How is HBM technically different from the DDR5 memory in most PCs? HBM stacks multiple DRAM dies vertically through TSVs across a very wide interface (up to 2048 bits with HBM4), versus DDR5's narrower interface; DDR5 launched at 4.8 Gbps per pin (jedec.org) versus HBM4's 2 TB/s aggregate bandwidth per stack (see Table 3).

Which companies supply HBM for AI chips, and who leads? Samsung Electronics, SK hynix, and Micron are the three merchant suppliers. Counterpoint Research credited SK hynix with 62% of HBM shipments as of Q2 2025 (news.skhynix.com), though Samsung's share rose through 2026 as its HBM4 qualification progressed (see Table 2).

Will the shortage affect consumer PC and laptop prices? This report does not make a forecast about consumer PC and laptop prices.

Table 3 below summarizes the JEDEC standard generations discussed throughout this report as a technical quick reference.

Standard	JEDEC document	Publication date	Peak bandwidth	Interface / channels
HBM2	JESD235A	Jan. 2016	Up to 256 GB/s per device (jedec.org)	1024-bit, 8 channels (jedec.org)
HBM3	JESD238	Jan. 2022	819 GB/s per device (6.4 Gb/s per pin) (jedec.org)	1024-bit, 16 channels, up to 12-high stacks (jedec.org)
HBM3E	JESD238B.01	Apr. 2025 (jedec.org)	Up to 1,180 GB/s per stack at 9.2 Gbps (Samsung) (semiconductor.samsung.com); 9.6 Gbps (SK hynix) (news.skhynix.com)	1024-bit, 16 channels, up to 12-layer stacks

Standard	JEDEC document	Publication date	Peak bandwidth	Interface / channels
HBM4	JESD270-4 / JESD270-4A	Apr. 2025 / Dec. 2025 (jedec.org)	Up to 2 TB/s per stack (8 Gb/s per pin) (jedec.org)	2048-bit, 32 channels, up to 16-high stacks (jedec.org)
DDR5 (comparison)	JESD79-5	Jul. 2020	4.8 Gbps initial data rate (jedec.org)	Double DDR4's bandwidth (jedec.org)

As *Table 3* shows, each HBM generation has roughly doubled aggregate per-stack bandwidth and channel count relative to its predecessor, a cadence that DDR5's single-generation, incremental improvement over DDR4 does not match, which is the specification-level reason HBM, not DDR5, has become the primary bottleneck resource for AI accelerator designers even though DDR5 pricing has also risen sharply as a secondary effect of capacity reallocation.
