David Soden

MARKET ASSESSMENT | AUGUST 2026

# Micron Sets Your Token Price, Not Nvidia

KV cache, LPDDR5X and an 85 percent gross margin: what long-context agentic inference costs once memory is the binding constraint, at 100k to 1M tokens

**David Soden**

Independent analyst and practitioner, enterprise automation and applied AI

davidsoden.com/market-assessment

REPORT DS-2026-6094 · 2026 · VERSION 1.0 · PUBLIC

## About this report

Every substantive claim in this report carries an evidence classification and, where applicable, a numbered source. Facts are separated from vendor claims, third-party estimates, the author's assessments, and untested hypotheses. Forward-looking sections use scenarios with observable tripwires rather than point forecasts. Stated assumptions are listed before the analysis begins.



Market Assessment reports are published and commissioned at **davidsoden.com/market-assessment**. This document is analysis, not investment, legal, or procurement advice.

## Scope and evidentiary standard

This assessment covers the memory hierarchy that serves decode-stage inference for agentic workloads at context lengths from 100,000 to 1,000,000 tokens: HBM3E and HBM4 on accelerator packages, LPDDR5X and GDDR7 as capacity tiers, conventional server DRAM, CXL-attached pools, and SSD-backed context storage. It covers the four questions a buyer has to answer before committing capital: how large the key-value cache actually is for the models being marketed for agent work, what Intel gave up by choosing LPDDR5X for Crescent Island, what memory repricing did to the cost floor, and whether vendor efficiency claims survive a restatement that charges for memory.

Three things about this category's published numbers need discounting before anything else. First, accelerator vendors publish throughput at operating points they do not fully specify: NVIDIA's figures for Qwen3.8-Flash-Next on GB300 NVL72 give tokens per second per GPU and tokens per second per user without stating batch size, context length, or power draw, so the two headline numbers cannot be jointly located on a curve without an assumption I have to supply. Second, Intel declined to publish Crescent Island's memory bandwidth, which is the single number that determines whether the product's economics work; every quantitative statement about that part in this document therefore rests on a third-party estimate and is labelled as one. Third, per-gigabyte memory pricing from commercial data sites contradicts itself across sources by a factor of six in the same quarter, and I refuse to build a dollar figure on it.

Where a claim rests on my own arithmetic rather than a published figure, the arithmetic is shown in the text so a reader can rerun it with different inputs. Model configuration details for the two Chinese hybrid-attention models are partly undisclosed; where I have assumed a value, the assumption is named at the point of use and carried into the assumptions section.

| Label | Definition | How to treat it |
|---|---|---|
| FACT | Verifiable in the cited primary source: a filing, press release, official documentation, or standards body announcement. | Reliable as stated. Check the source if the exact wording matters. |
| VENDOR CLAIM | Reported by a vendor about its own business or product. Self-measured and not audited by any third party. | The vendor said it. Directionally useful, not a measurement. Definitions of the metric are usually undisclosed. |
| THIRD-PARTY ESTIMATE | A research firm forecast or survey result. Methodology is proprietary and generally not reproducible. | Use for direction and order of magnitude only. Do not build a model on it. |
| ASSESSMENT | The author's reasoned judgment from the cited evidence. Falsifiable, and the reasoning is shown so it can be argued with. | This is analysis, not reporting. Disagree with it where your evidence differs. |
| HYPOTHESIS | A proposition offered for testing. Not currently supported by sufficient evidence to assert. | Treat as a research question, not a conclusion. Each one names what would confirm or refute it. |
| SCENARIO | A structured future state with a stated subjective probability. The probability is the author's judgment, not a calculated figure. | Use the leading indicators attached to each, not the probability. The indicators are observable; the probability is not. |

## Assumptions this analysis rests on

1. KV cache is stored at FP8 where the runtime supports it, and BF16 where it does not. GLM-5.3-Flash ships FP8 weights and vLLM documents FP8 KV cache on Blackwell with BF16 required on Hopper. If FP8 KV proves unstable in production and operators fall back to BF16, every per-token cache figure in this document doubles and the concurrency figures halve. The direction of the argument does not change; the magnitudes do.

2. GLM-5.3-Flash's eleven full-attention layers use an MLA latent geometry comparable to DeepSeek-V3, which stores 512 latent dimensions plus a 64-dimension rotary key component per token per layer. Z.ai has not published the latent rank. If the rank is materially larger, GLM's per-token cache rises proportionally and the gap to grouped-query attention narrows. The claim that the gap exists survives; the claim that it is roughly fiftyfold does not.

3. Qwen3.8-Flash-Next's twelve sparse-attention layers use four KV heads at 128 dimensions. Alibaba has published the layer split and the sparse block budget but not the head geometry. Where this assumption drives a number I have said so. The structural claim, that twelve of forty-eight layers carry a growing cache, comes from disclosed material and does not depend on the assumption.

4. Intel's Crescent Island memory bandwidth is between 1.2 and 1.8 TB/s, centred on the 1.5 TB/s implied by a 1280-bit LPDDR5X-9600 subsystem. This is Chips and Cheese's reconstruction

from board photographs and capacity arithmetic, not an Intel disclosure. If Intel ships above 2.5 TB/s the compute-balance finding in this document is wrong and the product is better than I assess it. If Intel ships below 1.2 TB/s it is worse.

5. Micron's cost per bit is roughly flat across the four quarters under discussion. This is the assumption behind the statement that current DRAM pricing sits several times above the level that would produce a normal-cycle gross margin. Micron is executing node transitions that lower cost per bit over time, so the true normalisation is less severe than the arithmetic implies. Treat the figure as an upper bound on the required price decline, not a forecast of one.

6. A GB300 NVL72 rack draws roughly 137 kW under load. NVIDIA does not publish a single rack figure; the 132 to 142 kW range comes from integrator and deployment sources. Every tokens-per-joule figure for that platform scales inversely with this number.

## The call

ASSESSMENT Memory, not compute, sets the floor on long-context agentic inference, and the floor moved because DRAM repriced, not because accelerators did. Buyers sizing capacity on tokens per watt are measuring the wrong thing: at agentic context lengths the binding limit is how many concurrent sessions fit in resident memory, and that is set by the model's attention architecture more than by the silicon. A capacity-tier accelerator is therefore only economic against a linear or sparse attention model. I would change position if conventional DRAM contract prices fell back toward historical margins.

*My judgment on the evidence assembled here, nothing more. There is data I can't see and data I may have missed. New evidence can move me, including to the opposite position, and I reserve the right to move.*

## Executive summary

The question buyers keep asking is which accelerator delivers the most tokens per watt. That question had a useful answer while inference was compute-bound. For a decode-stage agent holding several hundred thousand tokens of context, it no longer does, because the thing that runs out first is resident memory capacity, and the price of resident memory tripled over three quarters while accelerator list prices did not.

### FINDING 1 A single 1M-token session on a grouped-query model needs more memory than the model's weights.

> THIRD-PARTY ESTIMATE A Llama-3-70B-class model with eighty layers, eight KV heads and 128-dimension heads stores 320 KiB of cache per token at BF16. One million tokens is 328 GB, against 140 GB of BF16 weights. The cache is 2.3 times the model. Weights are paid once per instance; cache is paid per session per token, which is why the ratio is the wrong comparison and per-agent memory is the right one. [14]

### FINDING 2 The two Chinese hybrid models cut that figure by roughly fifty times, and they did it in the same week.

FACT GLM-5.3-Flash runs 34 linear-attention layers against 11 full-attention layers across 45 total. Qwen3.8-Flash-Next runs 36 Gated DeltaNet layers against 12 sparse-attention layers across 48. Both are 3:1. Z.ai states the architecture cuts KV cache size 4.44 times against the previous GLM-5.3 and attention compute 3.01 times. Convergence on the same ratio by two labs, independently, one week apart, is a design constraint asserting itself, not a coincidence. [9] [10] [12]

### FINDING 3 Intel's 480 GB of LPDDR5X buys capacity the card cannot feed, unless the model is sparse.

ASSESSMENT On the estimated 1.5 TB/s bandwidth and 1.3 PFLOP/s FP8, Crescent Island needs a decode batch near 433 to balance its matrix engines. After holding a 329 GB FP8 checkpoint, 151 GB of memory remains, which is 238 concurrent 100k-token sessions on GLM-5.3-Flash and about ten on a BF16 grouped-query model of half the size. The XMX array is provisioned for a batch the memory cannot hold, and the shortfall goes from 1.8x to 18x between 100k and 1M context. [5] [6] [11]

### FINDING 4 Intel is not publishing bandwidth because the number is the argument.

HYPOTHESIS Asked directly at Hot Chips, Intel said it was not disclosing bandwidth at this time. A figure near 1.5 TB/s sits beside 8 TB/s on a B300 and roughly 2 TB/s on a single HBM4 stack. Confirmation would be a general-availability spec sheet quoting bandwidth as a per-SKU range rather than a number, indicating the LPDDR5X supply and speed grade are not locked. Refutation would be Intel publishing a single figure above 2 TB/s before shipping. [5] [6] [13]

### FINDING 5 Micron's quarter was almost entirely price, and an 84.9 percent gross margin is what the buyer is funding.

FACT Fiscal Q3 2026, ended 28 May 2026: revenue $41.46 billion, up 73.7 percent sequentially; non-GAAP gross margin 84.9 percent, up ten points sequentially; DRAM average selling prices up in the low-60s percent sequentially against bit shipments up low single digits. Guidance for Q4 is $50.0 billion at roughly 86 percent gross margin. At that margin, cost of goods is 15 cents of every revenue dollar. [1] [2]

### FINDING 6 NVIDIA's own figures imply a concurrency of 80, and 80 sessions do not fit at 1M tokens.

ASSESSMENT NVIDIA reports over 16,000 tokens per second per GPU and over 200 tokens per second per user for Qwen3.8-Flash-Next on GB300 NVL72, at FP8, with no batch size, context length or power figure given. Sixteen thousand divided by two hundred is eighty concurrent sessions. On the assumed head geometry, eighty sessions at 1M tokens is roughly 983 GB of cache per GPU against 288 GB of HBM3E. The headline throughput therefore cannot be a 1M-token number. [7] [8]

### FINDING 7 The Alibaba speedups NVIDIA cites are measured at a 90 percent prefix-cache hit rate.

VENDOR CLAIM The 7.6x prefill and 4.9x decode figures are stated at 1M-token context with 90 percent of the prefix already cached. That is the warm path. A fresh agent session on a new repository or a new document set has a cold context and gets none of it. The benchmark is not

> wrong; it measures a real production condition. It is the condition a buyer sizing worst-case capacity must not plan against. [8] [9]

**FINDING 8 Per-gigabyte memory prices published by commercial data sites do not survive checking, so the dollar restatement cannot be completed honestly.**

> ASSESSMENT One widely cited source puts HBM3E at about $8.33 per GB in August 2026 while quoting LPDDR5X contract pricing at about $12.16 per GB in the preceding quarter and LPDDR5X spot at about $1.87 per GB. Those three numbers cannot all be right, and two of the three would invert Intel's cost argument. I have not found an auditable public series for either part, and I decline to compute a cost-per-token figure from these. [21]

## What a token of context costs in memory

The useful unit is bytes of cache per token, because it multiplies cleanly by context length and by concurrency, and because it is the only quantity that separates the architectures cleanly.

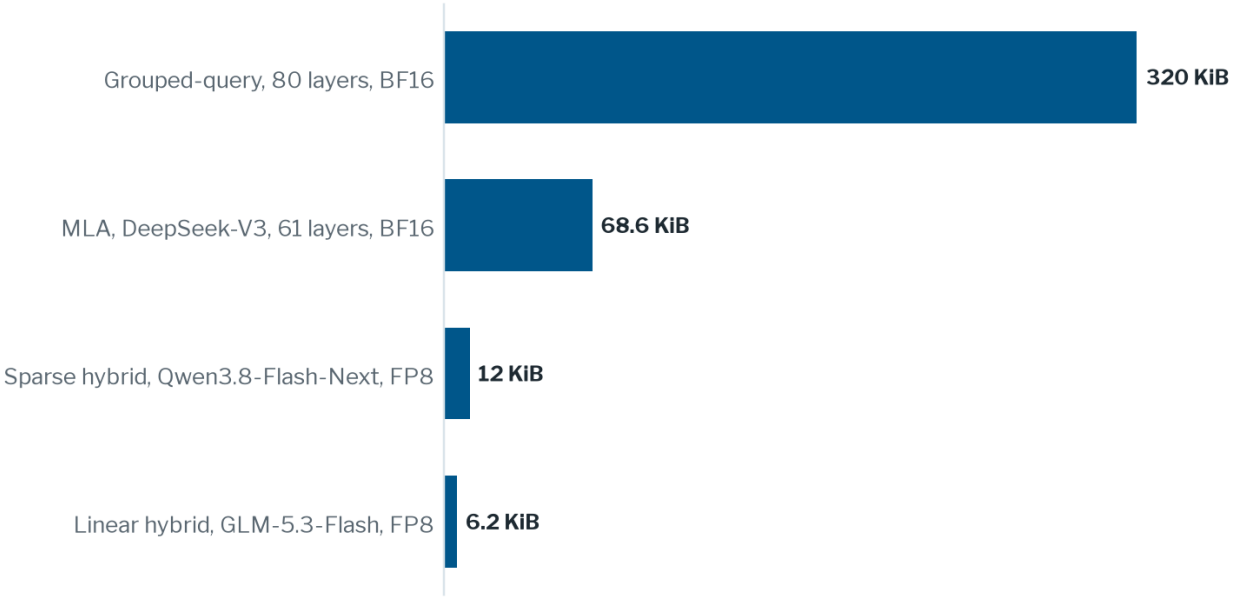
FACT Grouped-query attention stores two tensors, keys and values, for every layer and every KV head. For an eighty-layer model with eight KV heads at 128 dimensions in BF16, that is 2 x 80 x 8 x 128 x 2 bytes, or 327,680 bytes per token. Rounded, 320 KiB. [14]

FACT Multi-head latent attention replaces the per-head key and value tensors with a single compressed latent plus a small rotary component. DeepSeek-V3 stores 512 latent dimensions and 64 rotary dimensions across 61 layers, which at BF16 is 61 x 576 x 2, or 70,272 bytes per token. That is 68.6 KiB, a 4.7-fold reduction against the grouped-query figure above at the same precision. [14]

THIRD-PARTY ESTIMATE GLM-5.3-Flash carries a growing cache in only 11 of its 45 layers, the rest being Kimi Delta Attention layers whose state is fixed per sequence and does not grow with context. Assuming DeepSeek-comparable latent geometry and FP8 storage, that is 11 x 576 x 1, or 6,336 bytes per token: 6.2 KiB. Against the grouped-query baseline that is a factor of 52. The latent rank is the assumption; the layer split is disclosed. [9] [11]

THIRD-PARTY ESTIMATE Qwen3.8-Flash-Next carries a growing cache in 12 of 48 layers. Assuming four KV heads at 128 dimensions in FP8, that is 2 x 12 x 4 x 128, or 12,288 bytes per token: 12 KiB. The head geometry is the assumption. What is not an assumption is that three quarters of the layers are Gated DeltaNet and compress history into a fixed recurrent state, so the per-token growth applies to a quarter of the stack. [9] [12]

**FIGURE 1** **Bytes of KV cache per token of context, by attention architecture**




**THIRD-PARTY ESTIMATE** The first two figures follow directly from published layer counts and head geometry. The last two combine disclosed layer splits with assumed latent rank and head count, and mix precisions: the FP8 figures would double at BF16. Read this as an ordering with roughly correct spacing, not as four measurements. [9] [11] [12] [14]

## Sparse attention does not shrink the cache

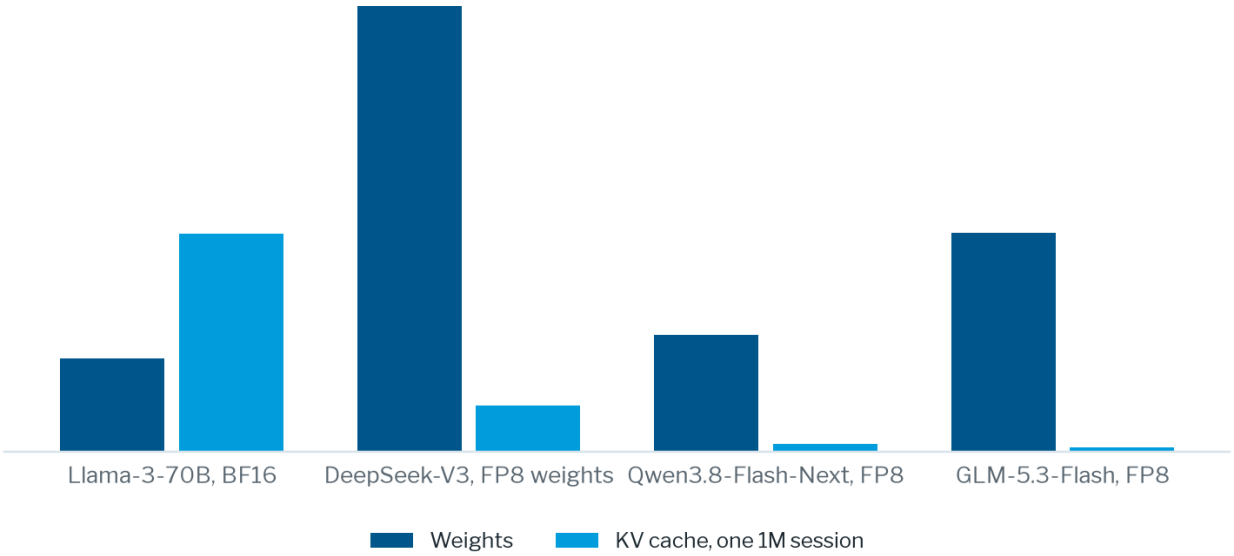
**ASSESSMENT** This is the point most often got wrong. GLM-5.3-Flash's full-attention layers use a top-2048 IndexPool and Qwen3.8-Flash-Next's use a top-512-block budget that works out to the same 2048 tokens. Both reduce how many cached tokens each query attends to, which cuts attention arithmetic. Neither reduces how many cached tokens must remain addressable, because the indexer has to be able to select any of them. Sparse attention is a compute optimisation with a memory bandwidth benefit and no capacity benefit. The capacity benefit in these models comes entirely from the linear layers. [9]

**ASSESSMENT** That distinction decides which memory tier the workload can tolerate. A compute optimisation lets you buy fewer FLOPS. A capacity optimisation lets you buy cheaper bytes. Only the second one makes an LPDDR5X part viable, and only the linear-attention layers deliver it.

## The comparison to weights is the wrong comparison

**ASSESSMENT** Weights are a fixed cost per served instance. Cache is a variable cost in two dimensions at once, sessions and tokens. Asking whether the cache exceeds the weights therefore has no fixed answer; it has a crossover concurrency. For GLM-5.3-Flash at FP8, weights of roughly 329 GB are matched by cache at about 519 concurrent 100k-token sessions, or 52 sessions at 1M. For the grouped-query model, the crossover at 1M context is below one: a single session already exceeds the weights. [11] [14]

**FIGURE 2** **Model weights against KV cache for one 1M-token session**




THIRD-PARTY ESTIMATE Weights are per instance and paid once. Cache is per session and scales with concurrency, so the right-hand bars multiply by the number of active agents while the left-hand bars do not. DeepSeek-V3 cache is shown at BF16 because that is the published figure; the other two cache bars are FP8. [11] [14]

ASSESSMENT For anyone sizing a deployment the operative number is not either bar. It is the free capacity after weights, divided by cache per session. That quantity is what the next section applies to a specific card.

## Intel's 480 gigabytes and the number it will not print

FACT Crescent Island, detailed at Hot Chips 2026, is a PCIe Gen5 x16 accelerator with 32 Xe3P cores feeding 256 XMX engines on a 16-deep systolic array, 1 MB of general register file and 512 KB of L1 and shared local memory per core, a 32 MB unified L2, and 160 GB of LPDDR5X as standard with configurations to 480 GB. Board power is 350 W, air-cooled, with active idle at or below 50 W and low-power idle near 10 W. [5] [22]

FACT Crescent Island was shown alongside Diamond Rapids and Wildcat Lake as part of a hardware set Intel presented for the next generation of agentic AI workloads, which is the framing the company chose for the launch rather than a claim about any part's performance. [18] [22]

FACT Intel did not disclose memory bandwidth. Asked at the session, the answer was that the company is not disclosing that information at this time. [5] [6]

THIRD-PARTY ESTIMATE Chips and Cheese reconstructs a 1280-bit LPDDR5X subsystem from board photography and capacity arithmetic, which at LPDDR5X-9600 gives over 1.5 TB/s, and estimates 1.3 PFLOP/s FP8 and 2.6 PFLOP/s at FP4 or MXFP4 at an assumed 2.5 GHz. Both figures are the author's reconstruction, not Intel's. [6]

| Part | Memory type | Capacity | Bandwidth | Board power |
|---|---|---|---|---|
| **Intel Crescent Island** | LPDDR5X | 160 to 480 GB | Undisclosed; est. 1.5 TB/s | 350 W |
| **NVIDIA B300, in GB300 NVL72** | HBM3E | 288 GB | 8 TB/s | up to 1,400 W |
| **NVIDIA Rubin CPX** | GDDR7 | 128 GB | approx. 2 TB/s | Not established here |
| **AMD MI350P** | HBM3E | Not established here | 3.6 to 4 TB/s | Not established here |
| **One HBM4 stack, JEDEC maximum** | HBM4 | Per JESD270-4 | up to 2 TB/s | n/a |

ASSESSMENT The row that should stop a reader is the last one. On the estimated figure, the entire 1280-bit LPDDR5X array on Crescent Island delivers roughly what one HBM4 stack delivers under the JEDEC maximum. A B300 carries eight stacks of the prior generation. [6] [13]

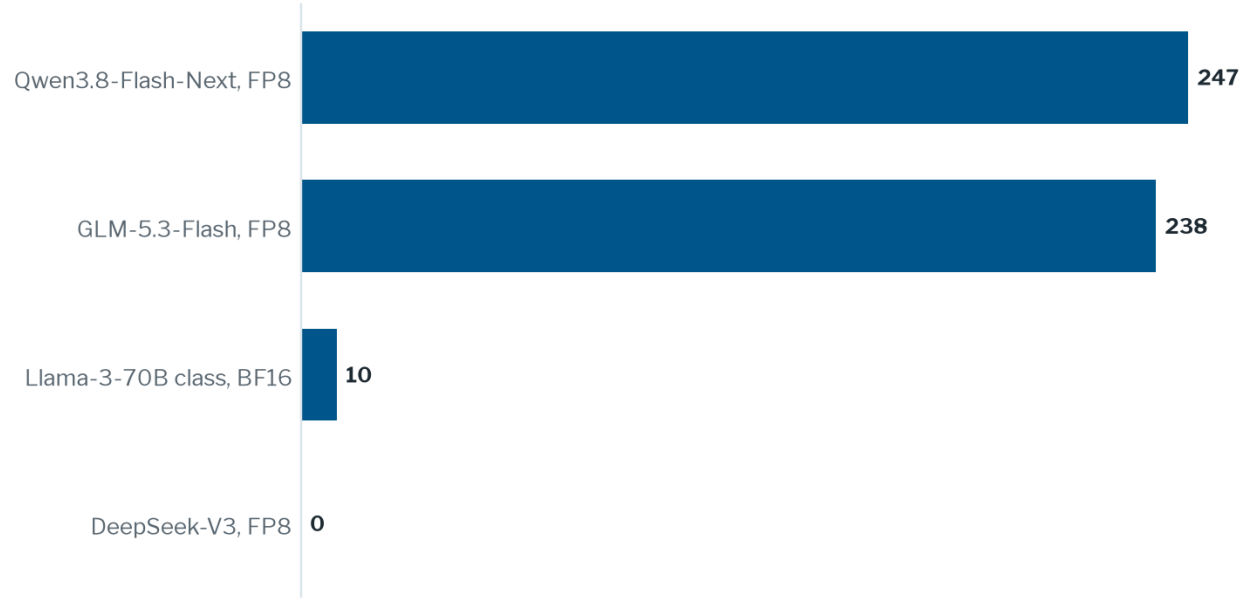
## What the missing number implies

ASSESSMENT Take the estimates at their centre. To keep 1.3 PFLOP/s busy against 1.5 TB/s requires an arithmetic intensity of about 867 FLOP per byte. Decode on an 18-billion-active-parameter model costs about two FLOP per active parameter per token per sequence, and reads the active weights once for the whole batch, so the batch that balances the machine is roughly half the intensity: about 433 concurrent sequences. [6] [9]

THIRD-PARTY ESTIMATE Capacity does not reach that batch. A 480 GB card holding GLM-5.3-Flash's roughly 329 GB FP8 checkpoint has 151 GB left. At 6,336 bytes per token, 100k context consumes 634 MB per session, giving 238 sessions. At 1M context it consumes 6.34 GB, giving 24. The card is short of its own compute balance by a factor of 1.8 at 100k and 18 at 1M. [6] [11]

THIRD-PARTY ESTIMATE Substitute a dense grouped-query model and the picture collapses. A 70-billion-parameter BF16 checkpoint leaves 340 GB free, and at 32.8 GB per 100k-token session that is ten concurrent agents on a 480 GB card. DeepSeek-V3 at FP8 does not fit at all, its weights alone exceeding the card. [14]

**FIGURE 3** **Concurrent 100k-token sessions in 480 GB, after model weights**




THIRD-PARTY ESTIMATE Free capacity after weights divided by cache per session. DeepSeek-V3 scores zero because its 671 GB FP8 checkpoint exceeds the card before any cache is allocated, not because its cache is large. Sensitive to the FP8 assumption: at BF16 cache the first two bars roughly halve. [11] [14]

ASSESSMENT Intel has built a card whose commercial case is coupled to an architectural choice made in Beijing and Hangzhou. That is a defensible bet, because the 3:1 convergence is evidence the choice is being forced by the same physics everywhere. It is also a bet that a Western enterprise procurement function has to be willing to buy an open-weight Chinese model to realise. Intel's marketing does not mention this, and it is the largest single risk to the product. [9]

## Why the bandwidth figure is missing

HYPOTHESIS Two explanations fit. The first is presentational: 1.5 TB/s printed next to 8 TB/s frames the product as a fifth of a B300 rather than as 1.7 times the capacity, and a launch does not want that headline. The second is supply: LPDDR5X speed grade and bus population are procurement decisions in a market where module contract prices moved sharply in the same window, and committing to a bandwidth figure commits to a bill of materials Intel may not be able to hold across SKUs. [5] [6]

HYPOTHESIS These are separable. If the general-availability specification quotes bandwidth as an up-to figure varying by SKU, the supply explanation is confirmed. If Intel publishes one number above 2 TB/s before shipping, both are refuted and the compute-balance finding above is wrong in Intel's favour. If Intel ships without ever publishing bandwidth, the presentational explanation is the live one and buyers should treat the omission as an admission.

ASSESSMENT There is a comparison that flatters Intel and should be stated. Bandwidth per watt at 1.5 TB/s and 350 W is about 4.3 GB/s per watt; a B300 at 8 TB/s and 1,400 W is about 5.7. Intel is not giving up much bandwidth efficiency. What it gives up is bandwidth density per socket, which converts into more cards, more PCIe fabric and more failure domains for the same

aggregate. That is a scale-out cost, not a power cost, and it lands on the buyer's rack budget rather than the buyer's electricity bill. [5] [6] [7]

**ASSESSMENT** One further reading. NVIDIA reached the same conclusion about memory tiering a year earlier and applied it in the opposite direction. Rubin CPX pairs 128 GB of GDDR7 at roughly 2 TB/s with 30 PFLOPS of NVFP4 and aims it at prefill, which is compute-bound. Intel has taken a comparable bandwidth tier and aimed it at decode, which is not. Whether that is insight or error is the question the undisclosed number would settle. [15]

## The cost floor moved, and it was not the GPU

**FACT** Micron reported fiscal Q3 2026 on 24 June 2026 for the quarter ended 28 May 2026: revenue $41.46 billion, up 73.7 percent sequentially and 345.6 percent year over year; GAAP gross margin 84.6 percent and non-GAAP 84.9 percent; GAAP operating income $33.32 billion; non-GAAP diluted EPS $25.11; operating cash flow $25.39 billion against net capital expenditure of $7.1 billion and adjusted free cash flow of $18.3 billion. [1]

**FACT** In the same reporting, DRAM revenue was $31.3 billion, or 76 percent of the total, with average selling prices up in the low-60s percentage range sequentially and bit shipments up in the low single digits. Non-GAAP gross margin rose ten percentage points sequentially. Guidance for fiscal Q4 is revenue of $50.0 billion plus or minus $1.0 billion at roughly 86 percent gross margin. [1] [2]

**ASSESSMENT** The decomposition is worth doing because it is the whole story. Prices up about 62 percent against bits up about 3 percent implies DRAM revenue up about 67 percent, which back-solves to roughly $18.8 billion in the prior quarter. Seventy-six percent of Micron's reported fiscal Q2 revenue of $23.86 billion is about $18.1 billion. The two agree within a few percent, so the decomposition holds: essentially none of the growth was volume. [1] [2]

**FIGURE 4** **Micron fiscal Q3 2026 DRAM growth: price against volume**


**FACT** Sequential percentage change. The price and bit figures are as characterised by Micron in its fiscal Q3 prepared remarks; the third bar is the product of the first two and is arithmetic, not a reported figure. Micron gave ranges, not points; midpoints are used. [1] [2]

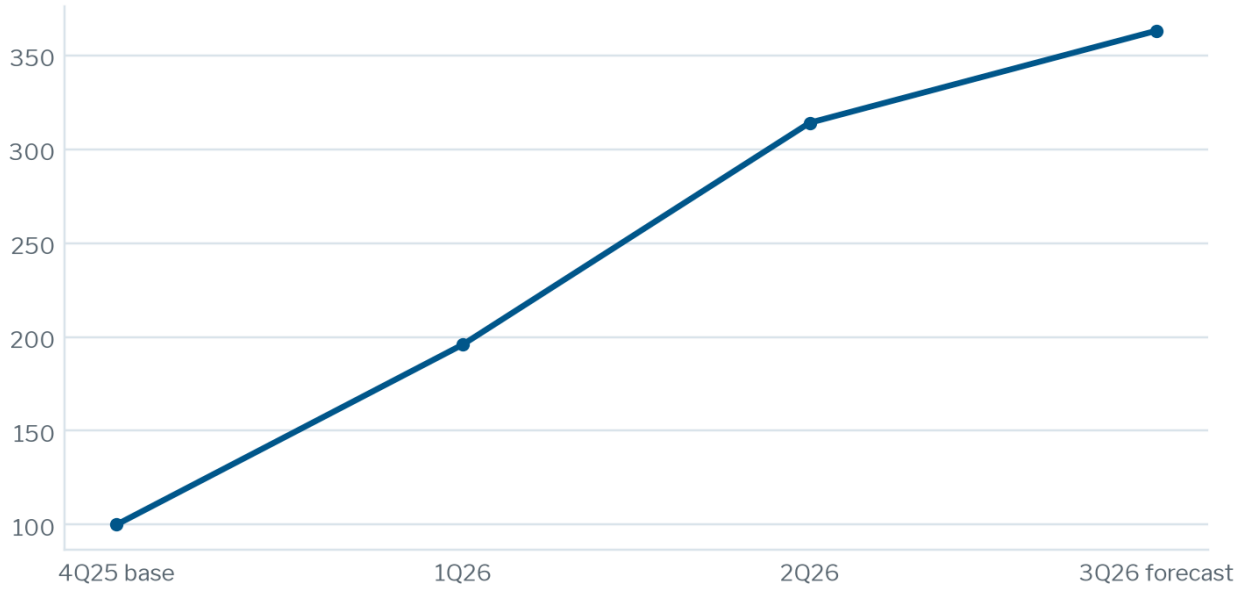
## What an 84.9 percent margin implies about the price you pay

**ASSESSMENT** At 84.9 percent non-GAAP gross margin, cost of goods sold is 15.1 cents of every revenue dollar. Hold unit costs and volumes constant and ask what revenue per bit would produce a 35 percent gross margin, which is unremarkable for a memory supplier across a full cycle. The answer is 15.1 divided by 0.65, or 23.2 percent of the current level: a decline of about 77 percent. Put the other way, today's DRAM price is roughly four and a third times the price that would produce an ordinary margin at today's cost base. [1]

**ASSESSMENT** Two corrections in opposite directions. The 84.9 percent figure is consolidated and includes lower-margin NAND, so the DRAM-only margin is higher and the implied overshoot larger. Against that, node transitions lower cost per bit over time, so the price that produces a 35 percent margin two years from now is lower than the price that produces it today, and the required decline is correspondingly smaller. Treat 77 percent as an upper bound on the correction, not a forecast of one. [1]

**THIRD-PARTY ESTIMATE** TrendForce's published sequence for conventional DRAM contract prices is a rise of 93 to 98 percent quarter on quarter in 1Q26, 58 to 63 percent in 2Q26, and a forecast 13 to 18 percent in 3Q26. Compounded at midpoints that is a factor of about 3.6 over three quarters. One of those three quarters is a forecast and one was published as an expectation rather than a print, so the settled portion of the move is closer to 3.1 times. [3] [4]

**FIGURE 5** **Conventional DRAM contract price index, compounded from TrendForce quarterly changes**

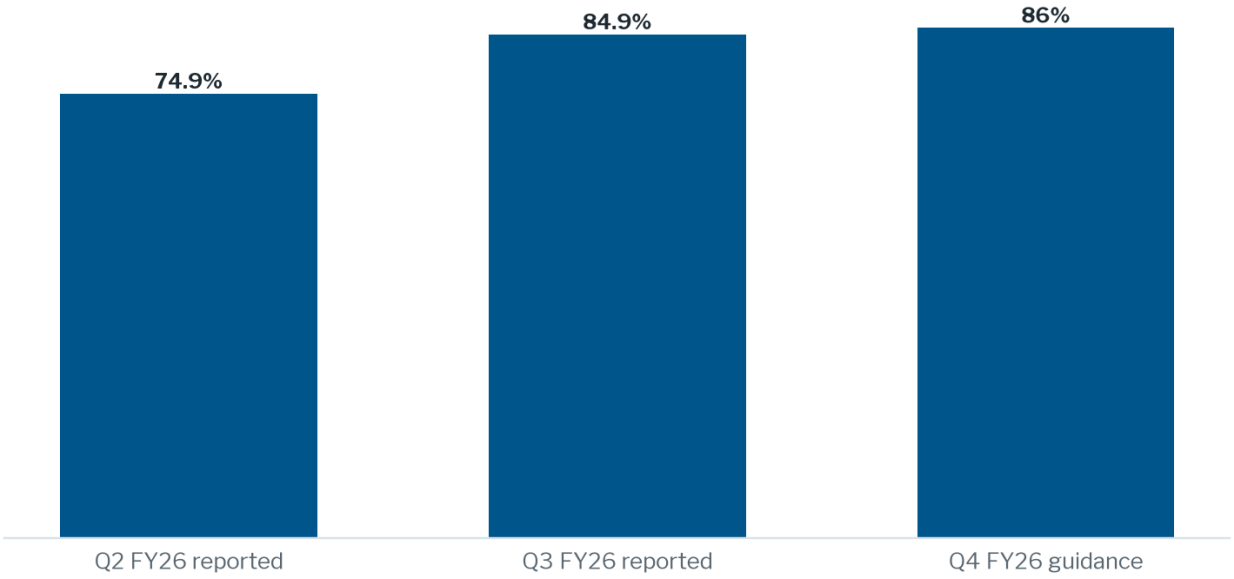



THIRD-PARTY ESTIMATE Index built by compounding the midpoints of TrendForce's published quarter-on-quarter ranges from an arbitrary base of 100. The final point is a forecast and the third was published as an expectation, so the line hardens as it moves left. This is an analyst firm series with proprietary methodology, not a settlement price. [3] [4]

FACT Micron has roughly $100 billion of contracted revenue under multi-year strategic customer agreements, supported by about $22 billion in cash deposits and letters of credit. Its chief executive framed those agreements as enhancing the durability and predictability of financial performance. [1] [2] [17]

ASSESSMENT That is the sentence a CFO should read twice. The current price level is not merely a spot condition a patient buyer can wait out; a substantial share of forward supply has been contracted, with cash down, at prices set during the move. Whoever signed those agreements has locked the memory component of their inference cost for the term. Whoever did not is buying residual supply in a market where the largest buyers have already been served. [1] [2]

**FIGURE 6** **Micron consolidated gross margin, reported and guided**




**FACT** Non-GAAP. The Q2 figure is derived from Micron's statement that Q3 margin rose ten percentage points sequentially. The third column is management guidance issued in the same release, not a result, and should be discounted accordingly. [1] [2]

### The per-gigabyte numbers do not survive checking

**ASSESSMENT** A memory-cost-inclusive restatement wants a price per gigabyte for HBM3E and LPDDR5X. The public sources for these contradict each other. One commercial data site puts HBM3E near $8.33 per GB in August 2026. A module-level figure for LPDDR5X contract pricing in the preceding quarter works out to about $12.16 per GB after a reported 89 percent quarterly rise. A spot figure for the same part type works out to about $1.87 per GB. These differ by more than sixfold, and two of the three would make LPDDR5X more expensive per gigabyte than HBM3E, which would invert Intel's entire cost argument. [21]

**ASSESSMENT** None of these sites publishes a methodology or a settlement source. I have not been able to reconcile them against any filing or contract disclosure, and I am not going to build a cost-per-token figure on them. The absence is itself the finding: the central quantitative claim behind capacity-tier accelerators, that LPDDR5X bytes are meaningfully cheaper than HBM bytes, is not verifiable from public 2026 data. Buyers should demand it in writing from the vendor, priced at the configuration they intend to purchase. [21]

## Tokens per watt, restated

**VENDOR CLAIM** NVIDIA reports Qwen3.8-Flash-Next running on GB300 NVL72 at over 16,000 tokens per second per GPU peak and over 200 tokens per second per user, at FP8, on 72 Blackwell Ultra GPUs with 130 TB/s of all-to-all NVLink bandwidth. Batch size, context length, sequence mix and power draw are not stated, and no tokens-per-watt figure is given. [8]

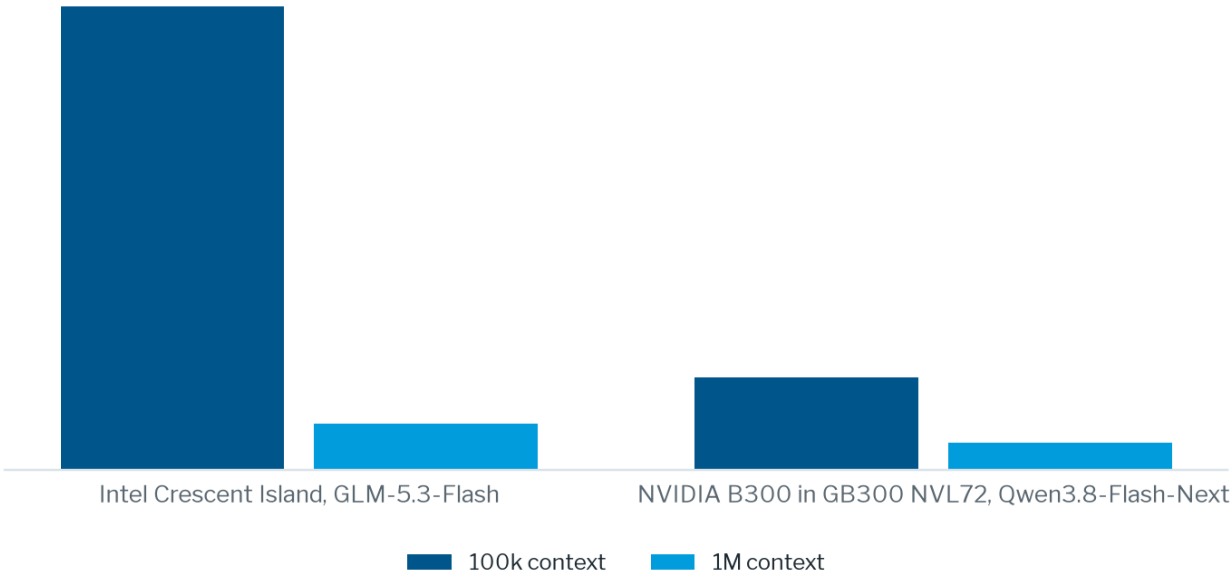
**ASSESSMENT** If those two figures describe the same operating point, concurrency is 16,000 divided by 200, or 80 sessions per GPU. NVIDIA does not say they describe the same point, and

a peak throughput number and an interactivity number are frequently taken from different points on a latency-throughput curve. Eighty is therefore an upper bound on the implied concurrency and the most favourable reading of the disclosure. [8]

THIRD-PARTY ESTIMATE Test that concurrency against memory. At the assumed 12,288 bytes per token, 80 sessions holding 100k tokens is about 98 GB of cache per GPU, which fits inside 288 GB of HBM3E alongside a sharded checkpoint. The same 80 sessions at 1M tokens is about 983 GB per GPU, which exceeds the package by more than three times. The headline number is therefore not a 1M-token number, and NVIDIA does not claim it is. [7] [8]

THIRD-PARTY ESTIMATE Capacity-constrain the same GPU at 1M context and concurrency falls to roughly 23 sessions. Holding 200 tokens per second per user, per-GPU throughput falls to about 4,600 tokens per second, a 3.5-fold reduction with no change to the silicon, the model, or the power draw. That is the memory-inclusive restatement in physical units: the efficiency loss between 100k and 1M context is a residency effect before it is a compute effect. [7] [8]

**FIGURE 7 Tokens per joule at two context lengths, on the models each part is marketed for**




THIRD-PARTY ESTIMATE Crescent Island figures rest on estimated bandwidth and FLOPS and an assumed cache geometry; at 1.2 to 1.8 TB/s the 100k figure ranges from about 45 to 68. NVIDIA figures divide the company's disclosed per-GPU throughput by a 1,400 W part TDP, which excludes CPU, NVLink switch and cooling overhead and so flatters the platform. Do not read the gap as a measured result. [6] [7] [8]

ASSESSMENT Two readings of that chart, and both matter. The first is that a 350 W air-cooled part plausibly beats a 1,400 W liquid-cooled one on tokens per joule at 100k context, which is the entire case for the capacity tier and appears to hold on this arithmetic. The second is that Intel's advantage erodes tenfold between 100k and 1M while NVIDIA's erodes three and a half fold, because the constraint that binds Intel is capacity and capacity is what runs out first. A buyer whose agents run at 100k should look hard at the capacity tier. A buyer whose agents run at 1M should not.

ASSESSMENT A result this large should make a reader suspicious, and I would rather name the falsifiers than defend it. It fails if Intel's real bandwidth is below about 1.2 TB/s, if the LPDDR5X array cannot sustain rated bandwidth inside a 350 W board budget while the XMX engines are also drawing, if FP8 KV cache proves unusable and the concurrency halves, or if GLM-5.3-Flash's latent rank is materially larger than DeepSeek's. Any one of those moves the left-hand bars down by a third or more.

VENDOR CLAIM The Alibaba speedups NVIDIA reproduces, 7.6 times on prefill and 4.9 times on decode against full attention at 1M context, are stated at a 90 percent prefix-cache hit rate. At that hit rate nine tenths of the prompt is already computed and resident, so the measurement is of the warm path. Agentic coding sessions that reopen the same repository do hit that path; sessions that open a new one do not. The figure is a real production condition and a poor capacity-planning assumption. [8] [9]

ASSESSMENT Independent work reported in 2026 claims tokens per watt halves for every doubling of context window. Between 100k and 1M that relationship would predict a tenfold fall, against the 3.5-fold capacity-only fall computed above. I have not been able to trace that claim to a primary paper with a stated method and I am not relying on it. I note it only because it points the same direction and is more pessimistic than my own figure, so my restatement is the conservative one. [23]

## Where the spill goes: CXL, SSD and the tiers below HBM

FACT The serving stack now treats KV cache as a storage hierarchy rather than a GPU allocation. GPU HBM is the top tier, host DRAM the second, local NVMe the third, and shared remote storage over NVMe-oF or RDMA the fourth. LMCache and NVIDIA's NIXL transfer library are the software that moves cache between them. [16] [20]

VENDOR CLAIM Samsung published measurements in June 2026 of LMCache backed by a CXL switch-based CMM-D memory pool, reporting performance close to conventional DRAM in multi-GPU tensor-parallel configurations. The figure circulating from that work is roughly 92 percent of DRAM performance. This is the vendor's own measurement of its own part, and I was not able to retrieve the paper's test configuration to check what workload and context length produced it. [19]

ASSESSMENT Tiering changes the shape of the constraint without removing it. Spilling cache to CXL or NVMe converts a hard capacity ceiling into a latency tax on cache misses, which is the right trade for multi-turn sessions with high reuse and the wrong trade for a decode loop that touches its whole context every token. Sparse attention helps here in a way it does not help on capacity: a top-2048 budget means each step reads a bounded slice, which is exactly the access pattern a tiered store can serve. That is the strongest technical argument for the hybrid architectures that this document has not otherwise made. [9] [20]

ASSESSMENT The commercial consequence is that the memory suppliers capture value across the whole stack rather than only at HBM. A buyer who tiers to host DRAM to escape HBM pricing is buying conventional DRAM in a quarter when conventional DRAM contract prices compounded roughly threefold. There is no tier in this hierarchy that is not repricing. [3] [4]

## Scenarios to 2028

Point forecasts fail in this category because the two variables that matter, memory contract pricing and model attention architecture, are both moving at quarterly frequency and are only loosely coupled. Conventional DRAM contract prices went from a 95 percent quarterly rise to a forecast 15 percent in two quarters. Two labs shipped convergent architectures one week apart. Anyone quoting a 2028 cost per million tokens is quoting a number with no method behind it. The probabilities below are the least reliable content in this document; the earliest visible signs are the most useful.

### SCENARIO A Price plateau 40 percent

Conventional DRAM contract increases fall below 5 percent quarter on quarter by 1Q27 as committed capacity lands and consumer demand stays weak.

**Consequence.** The cost floor stabilises well above the 2025 level rather than reverting. Buyers who signed multi-year agreements during the move overpay against spot for the term but keep supply.

**Earliest visible sign.** TrendForce guiding 4Q26 conventional DRAM below 10 percent QoQ, or a Micron quarter where bit growth exceeds ASP growth for the first time since 4Q25.

### SCENARIO B Architecture absorbs the shock 30 percent

Linear and sparse hybrids become the default for agentic serving outside China as well as inside it, and the 3:1 ratio holds across labs.

**Consequence.** Memory demand per agent falls by more than an order of magnitude, capacity-tier accelerators become genuinely viable, and memory unit demand grows more slowly than the agent count.

**Earliest visible sign.** A frontier Western lab shipping a production model with a disclosed linear-attention layer ratio above 50 percent, or a major serving framework defaulting to a linear-attention model for long-context agent work.

### SCENARIO C The move keeps compounding 20 percent

HBM4 wafer allocation continues to crowd out conventional DRAM capacity through 2027 and prices resume double-digit quarterly increases.

**Consequence.** Memory becomes the majority of inference bill of materials and published inference prices rise for the first time, breaking the assumption that token prices only fall.

**Earliest visible sign.** Micron guiding a fifth consecutive quarter of gross margin expansion, or any hyperscaler itemising memory as a separate cost line in a capital expenditure disclosure.

### SCENARIO D The capacity tier fails commercially 10 percent

Crescent Island and comparable LPDDR5X parts do not find volume buyers because the models whose architecture justifies them are Chinese open-weight releases that Western procurement will not deploy.

> **Consequence.** The inference market stays effectively HBM-only, the cost floor does not fall, and Intel's inference programme loses another cycle.
>
> **Earliest visible sign.** Intel reaching general availability without publishing memory bandwidth, or launching without a named model partner whose architecture makes the capacity arithmetic work.

## Leading indicators

These are observable without private information. Each is a specific event rather than a trend, so it either happens or it does not.

| If this happens | It means | Moves toward | Where to watch |
|---|---|---|---|
| **Intel publishes a single Crescent Island bandwidth figure above 2 TB/s** | The compute-balance shortfall in this document is wrong in Intel's favour | Scenario B | Intel Newsroom, product brief at general availability |
| **Crescent Island specifications quote bandwidth as an up-to range by SKU** | LPDDR5X speed grade is a procurement variable, not a design point | Scenario D | Intel product brief, OEM configuration guides |
| **A frontier US or European lab ships a model with over 50 percent linear-attention layers** | The 3:1 convergence is physics, not a regional design school | Scenario B | Model cards on Hugging Face, vLLM and SGLang recipe pages |
| **Micron reports a quarter with bit growth above ASP growth** | The price move has exhausted itself and volume is carrying revenue again | Scenario A | Micron quarterly press release and prepared remarks |
| **Micron guides a fifth consecutive quarter of gross margin expansion** | Supply remains tighter than the current price level assumes | Scenario C | Micron fiscal Q1 2027 guidance |
| **A hyperscaler discloses memory as a separate line in capital expenditure** | Memory has become material enough to require its own disclosure | Scenario C | 10-Q and 10-K filings, capital expenditure footnotes |
| **NVIDIA publishes throughput with batch size and context length stated** | The category is moving to comparable benchmark disclosure | Scenarios A and B | NVIDIA developer blog, MLPerf Inference submissions |
| **An auditable public price series appears for LPDDR5X and HBM3E per gigabyte** | The capacity-tier cost argument becomes checkable | Any | TrendForce and DRAMeXchange contract series, supplier filings |

# Implications

## For infrastructure buyers sizing inference capacity

**ASSESSMENT** Stop sizing on tokens per second and size on concurrent sessions at your actual context length. The calculation is free capacity after weights, divided by cache per session, and it is the number that determines whether a deployment serves ten agents or two hundred and forty. Ask every vendor for throughput at a stated batch size and a stated context length, and treat a benchmark that omits either as unusable. If your agents run near 100k tokens, price a capacity-tier part seriously. If they run near 1M, the capacity tier's advantage has largely gone and you are back to buying HBM.

**ASSESSMENT** Test the FP8 KV cache path before you commit to a capacity plan built on it. Every favourable number in this document doubles in memory terms if you fall back to BF16, and the fallback is a runtime and hardware-generation question rather than a choice. [11]

## For CFOs approving multi-year compute commitments

**ASSESSMENT** You are being asked to underwrite a memory price that sits roughly four times above the level consistent with an ordinary supplier margin. That is not an argument against signing; supply is genuinely constrained and about $100 billion of it has already been contracted with cash down. It is an argument for structuring the term so the memory component is not fixed for its whole length. Ask specifically whether the price you are quoted embeds a memory pass-through, and whether it reprices downward as well as upward. [1] [2]

**ASSESSMENT** The second question to ask internally is what context length your agent workloads actually use, because the answer moves your cost per token by three to ten times and nobody has priced it into the business case. A commitment sized on 100k-token sessions that meets 1M-token sessions in production is short of capacity by a factor no negotiated discount recovers.

## For semiconductor analysts covering memory suppliers

**ASSESSMENT** The bear case for the memory suppliers is not a demand collapse, it is the 3:1 layer ratio. If linear and sparse hybrids become the default serving architecture, memory demand per agent falls by more than an order of magnitude while agent count keeps growing, and the two curves cross somewhere. That crossing is the risk current gross margins are not discounting, and it is observable in model cards well before it appears in bit shipments. [9]

**ASSESSMENT** The near-term signal to track is the composition of Micron's revenue growth rather than its level. A quarter in which bit shipments grow faster than average selling prices is the first evidence the price move has done its work, and on the current decomposition that has not happened yet. Watch also whether HBM4 continues to consume conventional DRAM wafer capacity, because that substitution is what has kept the conventional tier tight and it is the mechanism scenario C depends on. [1] [2]

### For platform architects choosing context window limits

ASSESSMENT A context window limit is a memory budget wearing different clothes, and it is the highest-leverage cost lever you control. On the figures here, moving a service's cap from 1M to 100k tokens raises the concurrent sessions per accelerator roughly tenfold. That is a bigger effect than any accelerator choice in this document.

ASSESSMENT If you are choosing a model rather than a cap, the attention architecture matters more than parameter count for serving cost. A 320-billion-parameter linear hybrid holds fifty times less cache per token than a 70-billion-parameter grouped-query model, which means the larger model is the cheaper one to serve at long context. That inversion is new and most capacity planning has not caught up with it. [9] [14]

ASSESSMENT Design for prefix reuse deliberately rather than hoping for it. The vendor benchmarks that make long context look affordable assume a 90 percent prefix-cache hit rate, and the difference between a system that reaches that and one that does not is architectural: stable system prompts, stable tool definitions, session affinity in routing, and a tiered cache store behind the GPU. Those are choices you make, not conditions you inherit. [8] [20]

## What this document cannot answer

These require primary research, a supplier relationship, or a bench, and none of them is closed by anything published.

What LPDDR5X and HBM3E actually cost per gigabyte at contract volumes in 2026. Public sources disagree by more than sixfold and none publishes a method. Without this, the central premise of the capacity-tier argument is unverifiable and no honest cost-per-token figure can be built.

Crescent Island's real memory bandwidth, and whether the LPDDR5X array sustains its rated figure inside a 350 W board budget while the XMX engines are drawing. Every quantitative statement about that part in this document turns on it.

The latent rank of GLM-5.3-Flash's MLA layers and the KV head geometry of Qwen3.8-Flash-Next's sparse layers. Both are inspectable in the released configuration files by anyone willing to download them, which is the cheapest gap here to close and would move several figures from estimate to fact.

What NVIDIA's 16,000 tokens per second per GPU was measured at: batch size, context length, sequence-length distribution, prefix-cache hit rate, and rack power at that point. Without these the figure supports no comparison to anything.

The memory share of an accelerator's bill of materials, which no vendor discloses and which determines whether memory repricing passes through to accelerator prices or is absorbed in vendor margin.

Whether Western enterprise procurement will deploy Chinese open-weight models at production scale. The capacity-tier accelerator case largely depends on it and it is a governance question rather than a technical one.

## Half-life of this assessment

Decays within six months: every figure derived from Micron's fiscal Q3, the TrendForce price index, the fiscal Q4 guidance, and the specific throughput numbers NVIDIA has published for GB300 NVL72. Memory pricing is moving at quarterly frequency and the guidance will be superseded in September 2026.

Decays within twelve months: Crescent Island's specifications and the bandwidth question, which resolve at general availability; the concurrency arithmetic for GLM-5.3-Flash and Qwen3.8-Flash-Next, which changes with every model revision; and the claim that hybrid attention is a Chinese design choice, which one Western release would end.

Decays within twenty-four months: the position of HBM4 relative to the capacity tiers, and the scenario probabilities. HBM4E qualification and the next NVIDIA and AMD generations both land inside this window.

Architectural, and unlikely to decay: the distinction between compute optimisation and capacity optimisation in attention design, the fact that weights are a per-instance cost while cache is a per-session-per-token cost, and the rule that arithmetic intensity determines the batch size an accelerator needs to stay busy. Those hold regardless of who ships what.

## Limitations

This assessment was produced independently. I hold no position in and have no commercial relationship with Intel, NVIDIA, AMD, Micron, Samsung, SK hynix, Z.ai or Alibaba, and no party reviewed or funded this document.

The evidence base is unbalanced by design and by availability. The financial claims rest on a company earnings release and prepared remarks, which are primary and strong. The Intel claims rest on conference reporting and one independent reconstruction, because Intel withheld the determining specification. The model claims rest on a mix of vendor documentation, an independent architecture comparison, and a serving-framework recipe page, none of which is a technical report. Two of the sources supplied at the outset, the TechRadar article and the Samsung CXL white paper, could not be read in full at the time of writing, and are cited only for claims corroborated elsewhere.

Several of the most-cited numbers here are my own arithmetic on published inputs rather than measurements. The tokens-per-joule comparison in particular stacks an estimated bandwidth on an estimated FLOPS figure on an assumed cache geometry, and a reader should treat it as an ordering with a wide error bar rather than a result. I have named the falsifiers in the text rather than at the end, so they sit next to the claims they undermine.

No bench work was performed. Nothing here was measured on hardware, and the concurrency figures ignore memory fragmentation, paged-attention block overhead, activation memory, and the working set of the serving runtime itself, all of which reduce usable capacity below the arithmetic. Treat every session-count figure as an optimistic ceiling.

The commercial-data-site pricing discussed above is flagged as unverified throughout and no conclusion in this document rests on it.

## Sources


[1] Micron Technology, Inc. "Micron Technology, Inc. Reports Record Results for the Third Quarter of Fiscal 2026." 24 June 2026, for the quarter ended 28 May 2026. *Vendor primary* investors.micron.com

[2] Micron Technology, Inc. "Fiscal Q3 2026 Earnings Call Prepared Remarks." 24 June 2026. *Vendor primary* investors.micron.com

[3] TrendForce. "Long-Term Agreements Cap Price Increases; Server DRAM Contract Prices Expected to Rise 13-18% QoQ in 3Q26." 9 July 2026. *Analyst firm* trendforce.com

[4] TrendForce. "Rapid Contract Price Surge Drives 1Q26 DRAM Industry Up 81% QoQ." 1 June 2026. *Analyst firm* trendforce.com

[5] ServeTheHome. "Intel Crescent Island 160GB to 480GB LPDDR5X AI GPU at Hot Chips 2026." 24 August 2026. *Independent press* servethehome.com

[6] George Cozma. "Hot Chips 2026: Intel's Crescent Island." Chips and Cheese, August 2026. Bus width, speed grade, bandwidth and FLOPS figures are the author's reconstruction, not Intel disclosures. *Independent press* chipsandcheese.com

[7] NVIDIA. "GB300 NVL72" product page. Accessed August 2026. *Vendor primary* nvidia.com

[8] NVIDIA. "Experiment with Qwen3.8-Flash-Next 176B Model on NVIDIA GB300 NVL72 for Agentic Coding." NVIDIA Technical Blog, August 2026. *Vendor primary* developer.nvidia.com

[9] MarkTechPost. "GLM-5.3-Flash vs Qwen3.8-Flash-Next: Two Chinese AI Labs Independently Converge on the Same Model Architecture." 28 August 2026. *Independent press* marktechpost.com

[10] Z.ai. "GLM-5.3-Flash Overview." Z.AI developer documentation. Accessed August 2026. *Vendor primary* docs.z.ai

[11] vLLM. "zai-org/GLM-5.3-Flash." vLLM Recipes. Accessed August 2026. *Practitioner primary* recipes.vllm.ai

[12] Qwen Team, Alibaba Group. "Qwen3.8-Flash-Next." GitHub repository. Accessed August 2026. *Vendor primary* github.com

[13] JEDEC. "JEDEC and Industry Leaders Collaborate to Release JESD270-4 HBM4 Standard: Advancing Bandwidth, Efficiency, and Capacity for AI and HPC." *Standards body primary* jedec.org

[14] Sebastian Raschka. "KV Cache / Token (bf16)." LLM Architecture Gallery. Accessed August 2026. *Practitioner reference* sebastianraschka.com

[15] NVIDIA. "NVIDIA Unveils Rubin CPX: A New Class of GPU Designed for Massive-Context Inference." NVIDIA Newsroom, 9 September 2025. *Vendor primary* nvidianews.nvidia.com

[16] Simon Mugo. "The AI Memory Stack: What Investors Need to Know." Investing.com, via Yahoo Finance Canada, 29 August 2026. Cited for framing only; the article carries no quantitative market data. *Independent press* finance.yahoo.com

[17] 24/7 Wall St. "Broadcom vs Micron: Why They're So Strong as Capital Shifts to Inference and Agentic AI." 25 August 2026. Secondary summary of Micron's prepared remarks; figures corroborated against sources 1 and 2 before use. *Third-party comparison (commercial interest possible)* 247wallst.com

[18] TechRadar Pro. "Diamonds, crescents and wildcats: Intel shows off its hardware for the next generation of agentic AI workloads." August 2026. Article body could not be retrieved at time of writing; listed for completeness and not relied upon. *Independent press (body not retrievable)* techradar.com

[19] Samsung Semiconductor. "Optimizing KV Cache Offloading to CMM-D in a CXL Switch-based Memory Pool." White paper, June 2026. Test configuration not retrievable at time of writing. *Vendor primary* samsung.com

[20] Blocks & Files. "Nvidia and its partners' KV Cache extenders." 30 March 2026. *Independent press* blocksandfiles.com

[21] Silicon Analysts. "HBM Memory Pricing and Specifications (2026): Cost per Stack and per GB." Accessed August 2026. Methodology and settlement source not published; figures conflict with other commercial

series by more than sixfold and are treated as unverified throughout. *Third-party comparison (commercial interest, methodology undisclosed)* siliconanalysts.com

[22] TechPowerUp. "Intel Details Crescent Island Graphics: 32 Xe3P Cores, up to 480 GB LPDDR5X Memory." August 2026. *Independent press* techpowerup.com

[23] Claim that tokens per watt halves per doubling of context window, circulating in 2026 secondary coverage of energy-aware inference benchmarking. Primary paper and method not traced; noted in the text and explicitly not relied upon. *Second-hand (unverified)*

## Rights and permitted use