

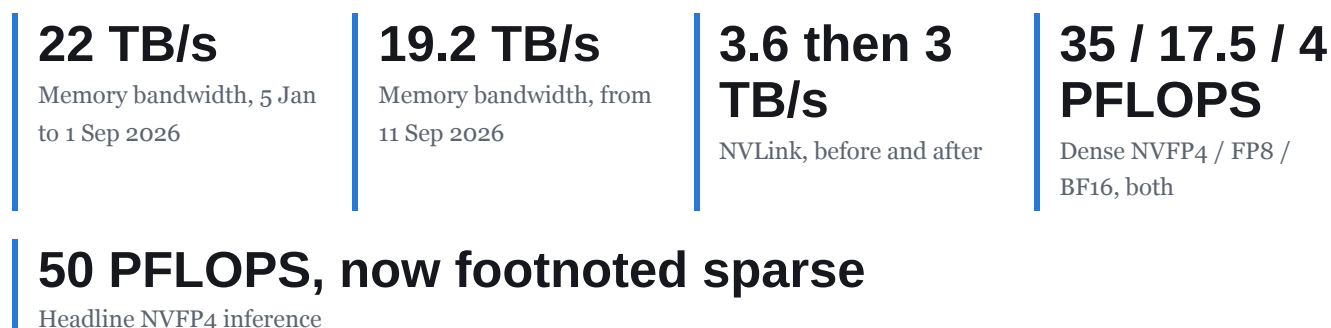
Sometime between 1 and 11 September 2026, NVIDIA edited the product page of Rubin, its newest accelerator. Its main figures for the arithmetic AI models use were left as they were. Its figure for memory bandwidth — how many bytes a second the chip can read from its own memory — fell from 22 terabytes a second to 19.2, the figure for its direct links to the other chips in a rack fell too, and the headline "50 petaFLOPS" acquired a footnote reading "Sparse specification" — a label NVIDIA's June datasheet already carried. The edit is small and unexplained, but it points straight at what an AI chip does and what limits it. The arithmetic is almost entirely matrix multiplication, which is why a processor built for graphics turned out to suit it. The limit that often binds, especially when a model is generating answers, is not arithmetic but memory: for a model writing one answer for one person, the multipliers on a current accelerator sit idle almost all of the time. Of what is advertised about these chips, how many there are is partly a matter of definition, and how many petaFLOPS each delivers is, for this kind of work, not the constraint.

1. The number that moved

NVIDIA's product page for the Vera Rubin NVL72, the rack-scale system built around Rubin, has been archived by the Internet Archive at least fifteen times this year. From the capture of 5 January 2026 to that of 1 September 2026, it lists each Rubin GPU with "288 GB HBM4 | 22 TB/s", above a footnote reading "Preliminary information. All values are up to and subject to change." In the capture of 11 September, and on the live page read on 24 September, the same row reads "288 GB HBM4 | 19.2 TB/s", a reduction of about 13%. The rack's aggregate memory bandwidth fell from 1,580 terabytes a second to 1,400, and the chip-to-chip NVLink figures fell further — from 3.6 terabytes a second per GPU to 3, and from 260 to 216 for the rack — while the scale-out networking figure rose, from 28.8 to 32.4 terabytes a second for the rack, now footnoted as bi-directional bandwidth. The "Preliminary information" note, which had been attached to the whole specifications table (rack, tray and GPU columns), is gone, and the table's rows keep only their sparse and dense footnotes; a new footnote, "Specifications based on an at-scale AI factory using NVIDIA DSX with MaxLPS" (a term the page does not define), is attached instead to a new table for a "100 MW" installation of "40K NVIDIA Rubin GPUs"; the inference row's headline figure of 50 petaFLOPS at NVIDIA's four-bit NVFP4 precision now carries a footnote reading "Sparse specification", bringing the page into line with the June datasheet, which already said "NVFP4 Inference specification is sparse". The dense four-, eight- and sixteen-bit figures — 35, 17.5 and 4 petaFLOPS per GPU — are unchanged; a few other rows, among them an INT8 figure and two emulated high-precision rows, were dropped. The revised page does not state one bandwidth throughout: 19.2 terabytes a second per GPU, 1,400 per 72-GPU rack (about 19.4 each) and "800 PB/s" for the 40,000-GPU installation (20 each).

The page does not explain the change, and other NVIDIA documents have not followed it. The preliminary Vera Rubin datasheet (PDF dated 22 June 2026), NVIDIA's HGX product page and a technical blog post of 21 July 2026 all still give 22 terabytes a second on 24 September. SemiAnalysis, an industry research firm, wrote on 25 February 2026 that "while Nvidia is targeting 22TB/s, we understand that memory suppliers are having challenges hitting Nvidia's requirements and we see it likely that initial shipments will come in slightly below at closer to 20TB/s". The page's 19.2 is a little below even that forecast; the direction matches, and NVIDIA has not said whether supply is the reason. NVIDIA said on 31 May 2026 that "production shipments of Vera Rubin are set to begin starting this fall", and on 26 August that the platform was "ramping into full production with racks running at partners including CoreWeave, Google Cloud, Microsoft Azure, Oracle Cloud Infrastructure and Nebius". MLCommons, which runs the MLPerf benchmarks, listed Rubin as "in preview" in results published on 16 September.

Figure 1. The Vera Rubin NVL72 product page, before and after September 2026, per Rubin GPU



Source: nvidia.com/en-us/data-center/vera-rubin-nvl72/ (rack NVLink 260 then 216 TB/s; a new footnote bases a new 100 MW table, not the per-GPU figures, on an at-scale AI factory using NVIDIA DSX with MaxLPS), Internet Archive captures of 5 January, 12 February, 16 and 25 March, 23 April, 28 May, 11 June, 11 and 28 July, 21 August and 1 September 2026 (22 TB/s, 'Preliminary information' footnote present) and of 11, 20 and 21 September 2026 (19.2 TB/s, footnote gone), plus the live page read 24 September 2026. Per-GPU values are the page's own; the rack figures are 1,580 then 1,400 TB/s. The exact day of the edit between 1 and 11 September is not recorded in any capture. NVIDIA's datasheet of 22 June 2026, its HGX page and its technical blog of 21 July 2026 still give 22 TB/s.

▼ Table view

Figure 1. The Vera Rubin NVL72 product page, before and after September 2026, per Rubin GPU

Measure	Value
Memory bandwidth, 5 Jan to 1 Sep 2026	22 TB/s
Memory bandwidth, from 11 Sep 2026	19.2 TB/s
NVLink, before and after	3.6 then 3 TB/s
Dense NVFP4 / FP8 / BF16, both	35 / 17.5 / 4 PFLOPS
Headline NVFP4 inference	50 PFLOPS, now footnoted sparse

The edit is a narrow event, but it isolates a variable. The numbers NVIDIA lowered were not Rubin's dense arithmetic ratings. They were how fast Rubin can be fed — from its memory, and from its neighbours.

2. Two readings that mislead

2.1 The count

At NVIDIA's GTC conference on 18 March 2025, Jensen Huang, the company's chief executive, used part of his keynote to revisit the name of the Blackwell generation. Tom's Hardware, reporting the keynote the same day, carried the point as a quotation:

Blackwell was named wrong.

The reasoning was physical. A Blackwell B200, sold as one GPU, contains two compute dies — separate pieces of silicon — in a single package. NVIDIA's own Blackwell datasheet says that "all NVIDIA Blackwell products feature two reticle-limited dies connected by a 10 TB/s chip-to-chip interconnect in a unified single GPU" — by the datasheet's own count, the pair is one GPU. The rack NVIDIA sold as the GB200 NVL72, named for its 72 packages, would on that logic have been an NVL144. SemiAnalysis, writing on 19 March 2025, summarised the new rule: "GPU counts are counted in terms of GPU dies in a package rather than the number of packages. This nomenclature will be adopted from Rubin onwards."

NVIDIA used the new unit in its own communications. A press release of 9 September 2025 refers to "customers looking to reuse existing Vera Rubin NVL144 systems".

Four months later the unit changed back, with little ceremony. NVIDIA's press release of 5 January 2026, issued at the Consumer Electronics Show, describes a "Vera Rubin NVL72" system "that combines 72 NVIDIA Rubin GPUs". The Register's Tobias Mann reported the same day that "it seems Nvidia has since changed its mind and is sticking with the established naming convention", and a caption beside his article says it will keep "counting SXM modules as GPUs rather than dies" — SXM being NVIDIA's name for the module that carries one package. The labels themselves outlived the rule. At GTC on 16 March 2026 NVIDIA's technical blog described a future "Vera Rubin Ultra NVL576" as eight racks "each with 72 Rubin Ultra GPUs", and a new rack design, Kyber, built "to fit 144 GPUs". For the NVL576, eight racks of 72 GPUs, the number counts packages; for Kyber, the blog's "144 GPUs" does not say which unit it counts, and Lockwood found the change of nomenclature made it hard to tell how much Kyber had shrunk. Glenn Lockwood, writing on his blog on 23 March 2026, observed that NVIDIA had "re-introduced the Rubin NVL72 nomenclature this year after making a point to rebrand Rubin NVL72 as Rubin NVL144 at last year's GTC", and that the change was made "without any commentary". NVIDIA's own releases do not give a reason.

Date	What changed	Where it is recorded
18 March 2025	Huang says Blackwell "was named wrong"; future products to be counted by die	Tom's Hardware (Jarred Walton), 18 March 2025; SemiAnalysis, 19 March 2025
9 September 2025	NVIDIA's newsroom uses "Vera Rubin NVL144"	NVIDIA press release, 9 September 2025
5 January 2026	NVIDIA's newsroom describes "Vera Rubin NVL72" with "72 NVIDIA Rubin GPUs"	NVIDIA press release; The Register (Tobias Mann), 5 January 2026

Date	What changed	Where it is recorded
16 March 2026	"NVL576" redefined as 8×72 packages; a Kyber rack to hold "144 GPUs"	NVIDIA Technical Blog, 16 March 2026
22 June 2026	Preliminary Rubin datasheet: "72 NVIDIA Rubin GPUs" per NVL72 rack	NVIDIA Vera Rubin datasheet (PDF dated 22 June 2026)
26 August 2026	"Vera Rubin, now in full production"	NVIDIA second-quarter fiscal 2027 results

A GPU count is the first misleading reading of a chip headline. The two-die design was the same throughout; what changed was the answer to a question that sounds trivial — how many of these things are there? — and the fact that it could be answered two ways, by the same company, within a year, shows how little a count of GPUs measures on its own.

"Ten thousand GPUs" suggests a quantity of some standard thing. In NVIDIA's Blackwell and Rubin products the thing is two dies in one package. Google's documentation for its newest tensor processing unit, TPU7x or Ironwood, notes that "frameworks like JAX expose each Ironwood chip as two separate 'devices'", one for each of its two chiplets. The same piece of hardware can therefore be one chip, two dies or two devices depending on who is describing it.

2.2 The headline speed

NVIDIA's H100 product page, read on 24 September 2026, lists the chip's eight-bit (FP8) performance as 3,958 teraFLOPS — almost four quadrillion floating-point operations a second. A footnote marks the figure "with sparsity". NVIDIA's Blackwell datasheet states the convention explicitly: "Specifications in sparse. Dense is one-half of the sparse spec shown." The H100's dense FP8 figure is therefore 1,979 teraFLOPS, and its dense sixteen-bit (BF16) figure 989. The larger number describes a special case; the smaller one describes ordinary work.

Even the dense figure does not say how quickly a model responds. For a single user waiting on a single answer, the question is not how fast the chip can multiply but how fast it can be fed, and for that workload an H100 could do its arithmetic roughly three hundred times faster than its memory can deliver the weights the arithmetic needs.

Figure 2. One chip, four headline numbers: the NVIDIA H100 SXM

Source: NVIDIA's H100 product page, specifications table, read 24 September 2026. The page prints sparse figures, marked with an asterisk footnoted 'With sparsity'; the dense figures are half, the convention NVIDIA's Blackwell datasheet (PDF dated 27 October 2025) states as 'Dense is one-half of the sparse spec shown'. The page lists BF16 at 1,979 TFLOPS with sparsity. All four are the vendor's peak figures; sustained rates in real programs are lower.

▼ Table view

Figure 2. One chip, four headline numbers: the NVIDIA H100 SXM

Measure	Value
FP8, as printed (with sparsity)	3,958 TFLOPS
FP8, dense	1,979 TFLOPS
BF16, dense	989 TFLOPS
Memory bandwidth	3.35 TB/s

3. The idea: a calculator at the end of a road

3.1 What the chip computes

A language model's parameters are organised as grids of numbers, one or more per layer. Each token the model processes is represented as a list of numbers, and passing that list through a layer means multiplying it by the layer's grid: each number is multiplied by a weight, and the products are added up, for every row of the grid. That operation is matrix multiplication, and it accounts for the bulk of the arithmetic a model performs. Jared Kaplan and colleagues at OpenAI, in their January 2020 paper on scaling laws, estimate the arithmetic of a model's forward pass at roughly two operations per parameter per token — "2N", plus a term that grows with the length of the context — and explain the factor:

the factor of two comes from the multiply-accumulate operation used in matrix multiplication

Applied to GLM-5.3, an open-weights model from the laboratory Z.ai which Artificial Analysis lists with about 40 billion of its parameters active on each token, that is roughly 80 billion operations for every token processed.

3.2 Why a graphics chip

Drawing an image on a screen involves much the same small calculation repeated for millions of pixels, most of which do not depend on one another. Matrix multiplication has the same structure: every entry in the result can be computed independently and simultaneously. Researchers noticed the fit well before the

current boom. In 2004 Ian Buck and colleagues at Stanford published *Brook for GPUs*, "a system for general-purpose computation on programmable graphics hardware" that used "the GPU as a streaming coprocessor", and reported programs running "up to seven times faster than their CPU counterparts". The paper already leaned on the quantity that turns out to matter most, citing William Dally and colleagues on arithmetic intensity, "the ratio of arithmetic operations to memory bandwidth", and defining a similar "computational intensity" of its own.

The application to neural networks followed. At the International Conference on Machine Learning in June 2009, Rajat Raina, Anand Madhavan and Andrew Ng of Stanford reported that their graphics-processor implementation of a deep network "is up to 70 times faster than a dual-core CPU implementation for large models", cutting the training of a network with 100 million parameters "from several weeks to around a single day". In 2012 Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton trained the image-recognition network whose variant won that year's ImageNet competition on "two GTX 580 3GB GPUs", over five to six days; they wrote that the network's size was "limited mainly by the amount of memory available on current GPUs and by the amount of training time that we are willing to tolerate".

By 2017 the chip had begun to be redesigned around the workload. NVIDIA's whitepaper for its Volta architecture introduced tensor cores, circuits dedicated to one operation: "Each Tensor Core operates on a 4x4 matrix", multiplying two small matrices and adding a third. On the V100, the first Volta chip, NVIDIA's datasheet lists 640 tensor cores and 125 teraFLOPS of "tensor performance", against 15.7 teraFLOPS of single-precision arithmetic from its 5,120 conventional cores. The eightfold advantage came from specialised matrix units working at lower precision, not from the number of cores.

3.3 The wall, and the roofline

The same 2009 Stanford paper contains the observation that matters most for the present. Its authors measured the multiplication of two 1,000-by-1,000 matrices on their graphics card: roughly 20 milliseconds in total, of which

the actual computation takes only 0.5% of that time

with the remainder "being used for transfer in and out of global memory". That transfer was between the computer's main memory and the card, a different path from the memory mounted on a modern accelerator's own package, and the paper itself says that "GPU computation and within-GPU memory accesses themselves are highly parallel"; the bottleneck it measured was the link to the card. What carries over is the shape of the finding: arithmetic had become cheap, and moving the numbers had not.

The principle had been stated earlier for processors in general. In a short note dated December 1994 and published in *ACM SIGARCH Computer Architecture News* in March 1995, William Wulf and Sally McKee of the University of Virginia observed that microprocessor speed was improving at a faster exponential rate than the speed of the DRAM that feeds it, and drew the conclusion:

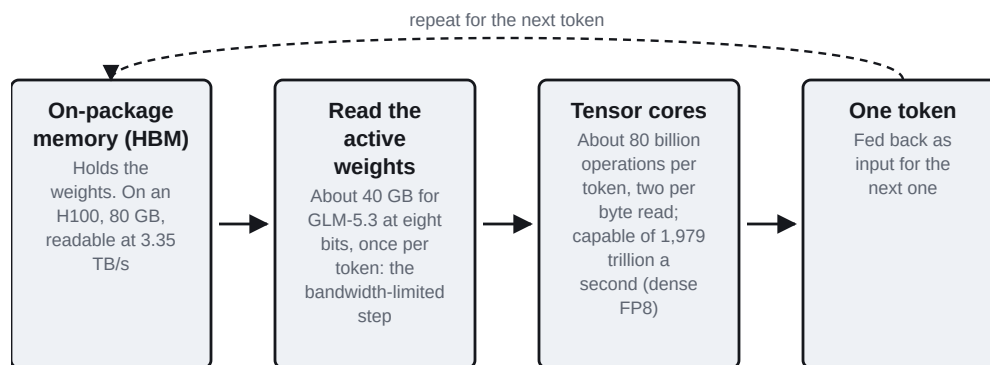
In fact, it will hit a wall.

Their model concerned latency — the average time a memory access takes — rather than bandwidth. McKee's own retrospective of 2004 notes that the bandwidth warning came earlier still, quoting John Ousterhout in 1989: "If memory bandwidth does not improve dramatically in future machines, some classes of applications may be limited by memory performance."

The version now used to reason about AI hardware came from Berkeley. In a technical report dated 17 October 2008, published in *Communications of the ACM* in April 2009, Samuel Williams, Andrew Waterman and David Patterson proposed a visual model they called the roofline, beginning from a premise:

We believe that for the recent past and foreseeable future, off-chip memory bandwidth will often be the constraining resource

The model reduces any program to one number, its "operational intensity", which the authors define as "operations per byte of DRAM traffic": how much arithmetic it performs for each byte it must fetch from main memory. They chose the term "instead of the terms arithmetic intensity" or machine balance because they wanted to count traffic between the caches and main memory, not between the processor and its cache; in wider use the looser "arithmetic intensity" is common. A chip's attainable performance is then the lesser of two limits: its peak arithmetic rate, or its memory bandwidth multiplied by the program's intensity. The crossover is the "ridge point", whose x-coordinate, in the report's words, "is the minimum operational intensity required to achieve maximum performance". A program to the left of the ridge is limited by memory; one to the right, by arithmetic.

Figure 3. Writing one token for one user: why the road, not the calculator, sets the pace

Schematic of generation for a single request (batch size one). The weights must all be read for each token; the arithmetic performed on them is about two operations per byte at eight-bit precision, against an H100 ridge point of about 591 operations per byte at FP8 (295 at BF16). Memory figures from NVIDIA's H100 product page and parameter counts from Artificial Analysis's GLM-5.3 page, both read 24 September 2026; operations per parameter from Kaplan et al., arXiv 2001.08361 (2020). Omitted: the attention cache, activations, and the fact that the full model does not fit on one H100.

▼ Table view

Figure 3. Writing one token for one user: why the road, not the calculator, sets the pace — stages

#	Stage	Note
1	On-package memory (HBM)	Holds the weights. On an H100, 80 GB, readable at 3.35 TB/s
2	Read the active weights	About 40 GB for GLM-5.3 at eight bits, once per token: the bandwidth-limited step
3	Tensor cores	About 80 billion operations per token, two per byte read; capable of 1,979 trillion a second (dense FP8)
4	One token	Fed back as input for the next one

Figure 3. Writing one token for one user: why the road, not the calculator, sets the pace — connections

From	To	Label
On-package memory (HBM)	Read the active weights	
Read the active weights	Tensor cores	
Tensor cores	One token	
One token	On-package memory (HBM)	repeat for the next token

3.4 Where a language model sits on the roofline

At sixteen-bit precision the ridge points of current accelerators, computed from their datasheets, lie between roughly 150 and 320 operations per byte; at eight bits they are roughly double, and at four bits they run from roughly 1,200 (B200, MI355X) to roughly 1,900 (B300). The four multicore machines in the 2008 Berkeley report had ridge points between 0.33 and 6.7 — double-precision arithmetic against measured bandwidth, where the modern figures are lower-precision vendor peaks, so the comparison overstates the change, though not its direction. Arithmetic has become so abundant relative to memory traffic that a program now needs to perform hundreds of operations on every byte it fetches in order to keep the multipliers busy.

A language model generating text for a single user is nowhere near that. To produce each token it must read every active weight from memory once, and it uses each weight for roughly two operations — a multiplication and an addition. With weights stored at eight bits, one byte each, that is about two operations per byte. On the H100 the FP8 ridge point is about 591.

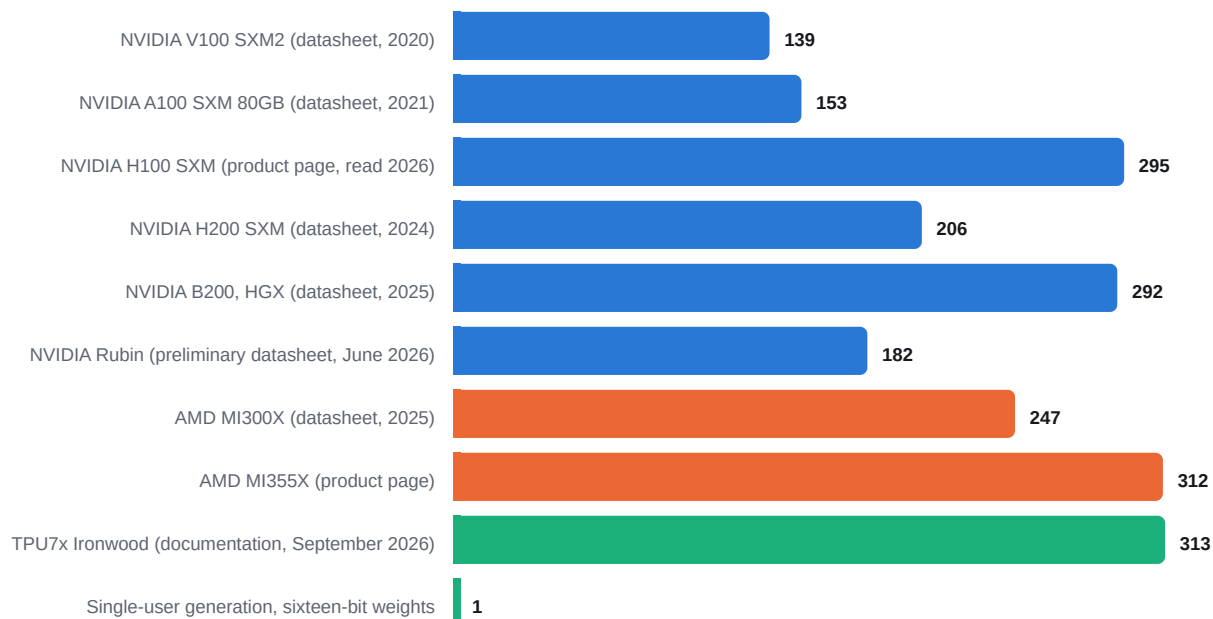
The consequence can be written as a single division. GLM-5.3's roughly 40 billion active parameters occupy about 40 gigabytes at eight bits. An H100 reads its memory at 3.35 terabytes a second. At most, then, it can stream those weights about 84 times a second, which caps single-user generation at about 84 tokens a second. Its dense FP8 arithmetic could in principle keep pace with roughly 24,700 tokens a second, about 300 times as many. For this workload the multipliers are idle more than 99% of the time.

The same division explains what a sparse, mixture-of-experts model does and does not save. GLM-5.3 holds roughly 753 billion parameters, all of which must be stored on the serving hardware because any of them may be selected for the next token; that is a question of memory capacity, and sparsity saves nothing on it. But to produce one token for one user it reads only the roughly 40 billion that run — about 40 gigabytes rather than about 750 — so on per-token memory traffic, the quantity that sets that ceiling, sparsity saves a great deal. "Memory" in these arguments names three different things: capacity (how much can be held), bandwidth (how fast it can be read) and latency (how long a single access takes, which is what Wulf and McKee modelled). They improve at different rates and limit different workloads.

The bound is a teaching bound rather than a prediction, and it omits several things that matter in practice. The whole of GLM-5.3 — roughly 750 gigabytes of weights if each is stored in eight bits — does not fit in one H100's 80 gigabytes, so a real deployment spreads it across ten or more chips, which then have to exchange results. Serving many users at once lets each weight fetched from memory be used for many tokens, which moves the workload towards the ridge. The attention cache adds memory traffic of its own; peak bandwidth is higher than what programs sustain, a gap McKee's 2004 retrospective already noted in the STREAM benchmark's measurement of "bandwidth sustainable by ordinary user programs, and not the theoretical 'peak bandwidth' that vendors advertise"; and some serving techniques emit more than one token per pass over the weights. None of these changes the direction of the argument. Amir Gholami and colleagues at Berkeley, in a 2024 study of the problem, conclude that memory bandwidth "can become the dominant bottleneck for decoder models", the family that includes today's chat models. NVIDIA's technical blog on the Rubin design, published on 21 July 2026, puts it more flatly, and then qualifies it:

The decode, or generation, phase of inference is fundamentally memory subsystem bound.

The next sentence adds: "It is not about peak bandwidth specs but rather how efficiently each kernel can utilize the entire memory subsystem." Bandwidth ceilings built from peak figures are therefore upper bounds twice over.

Figure 4. Break-even arithmetic per byte (ridge point), sixteen-bit dense, by accelerator

Ridge point = peak dense BF16/FP16 operations per second divided by peak memory bandwidth, each from the vendor's own document: V100 125 TFLOPS / 0.9 TB/s; A100 312 / 2.039; H100 989 / 3.35; H200 989 / 4.8; B200 (HGX) 2,250 / 7.7; Rubin 4,000 / 22 (preliminary; NVIDIA's product page gives 19.2 TB/s, which would put the ridge at 208); MI300X 1,307.4 / 5.3; MI355X 2,500 / 8; TPU7x 2,307 / 7.38. At eight-bit precision the ridge points roughly double on most chips (H100 591; Rubin 795). The last row is the intensity of generating one token for one user with sixteen-bit weights: about one operation per byte of weights read (two at eight bits, against eight-bit ridge points roughly double those shown). Vendor peak figures throughout; no independent body publishes competing peaks.

▼ Table view

Figure 4. Break-even arithmetic per byte (ridge point), sixteen-bit dense, by accelerator

Accelerator (source document)	Operations per byte of memory traffic at the ridge
NVIDIA V100 SXM2 (datasheet, 2020)	139
NVIDIA A100 SXM 80GB (datasheet, 2021)	153
NVIDIA H100 SXM (product page, read 2026)	295
NVIDIA H200 SXM (datasheet, 2024)	206
NVIDIA B200, HGX (datasheet, 2025)	292
NVIDIA Rubin (preliminary datasheet, June 2026)	182
AMD MI300X (datasheet, 2025)	247
AMD MI355X (product page)	312
Google TPU7x Ironwood (documentation, September 2026)	313
Single-user generation, sixteen-bit weights	1

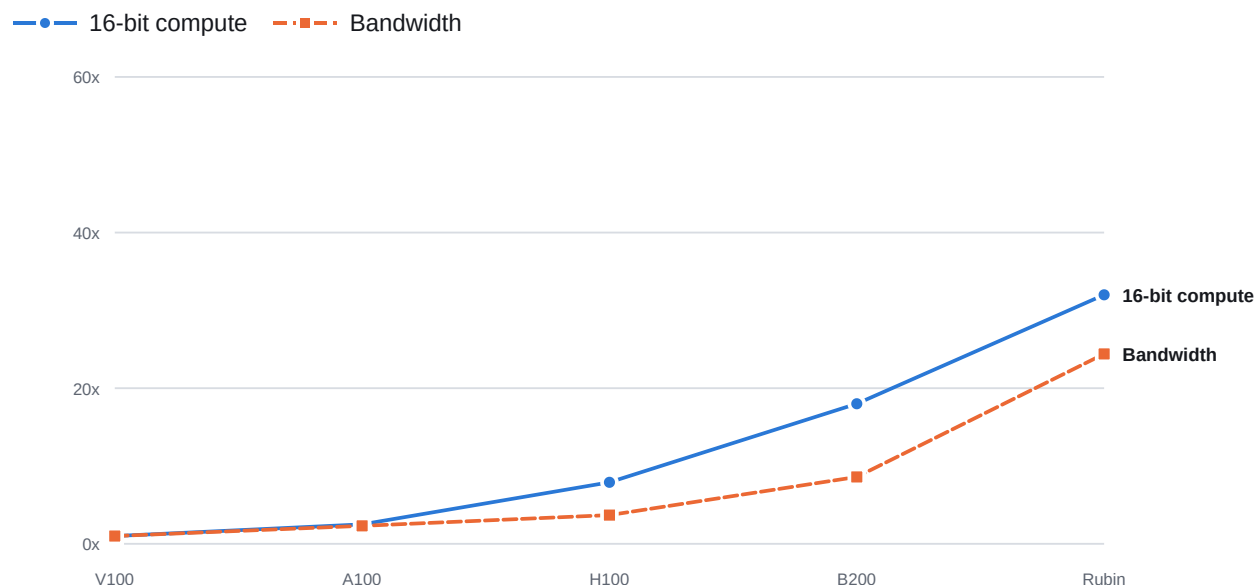
4. What the last nine years actually show

Put NVIDIA's flagship datasheets in sequence and, from 2017 to 2026 at least, the familiar story of arithmetic racing ahead of memory turns out to be largely a story about the size of the numbers.

NVIDIA's V100, introduced with the Volta architecture in 2017, is rated at 125 teraFLOPS of sixteen-bit tensor arithmetic and 900 gigabytes a second of memory bandwidth. Rubin, which NVIDIA said on 26 August 2026 was "now in full production", is rated on the company's preliminary datasheet of 22 June 2026 at 4 petaFLOPS of dense sixteen-bit arithmetic and 22 terabytes a second of bandwidth. At the same precision, compute grew about 32-fold and bandwidth about 24-fold. The gap between the two is real but modest.

The dramatic multiples come from lowering precision. Rubin's dense four-bit (NVFP4) training rating is 35 petaFLOPS, 280 times the V100's sixteen-bit figure — about 32 times from faster sixteen-bit arithmetic and about 9 from the smaller format. Nearly all of compute's lead over bandwidth comes from the format. Gholami and colleagues, in *AI and Memory Wall* (arXiv, March 2024; published in *IEEE Micro*), found that "over the past 20 years, peak server hardware FLOPS has been scaling at $3.0\times/2\text{yrs}$, outpacing the growth of DRAM and interconnect bandwidth, which have only scaled at 1.6 and 1.4 times every 2 years, respectively". The same paper records that the switch from 32-bit to 16-bit arithmetic "has enabled more than a $10\times$ increase in hardware compute capability", and its baseline is a 1990s processor, so its headline rates fold changes of precision and of architecture together. Between the H100 and Rubin, at sixteen bits, bandwidth grew faster than arithmetic (about 6.6 times against 4), and the sixteen-bit ridge point fell. Lower precision helps on both sides of the ledger: a weight stored in four bits takes a quarter of the memory traffic of one stored in sixteen, which is why falling precision has eased the bandwidth problem as well as inflating the arithmetic headline.

Figure 5. NVIDIA flagship accelerators at fixed precision: sixteen-bit compute and memory bandwidth (V100 = 1)



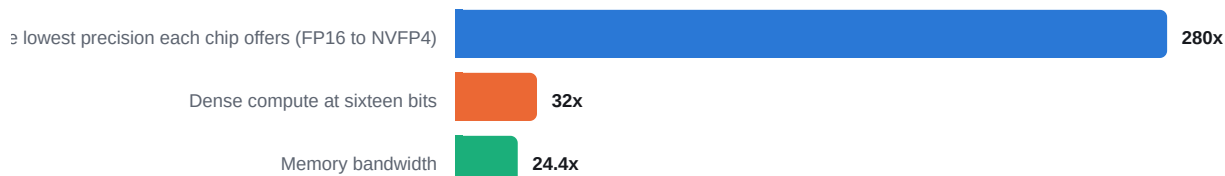
Computed from NVIDIA's own documents, read 24 September 2026. Sixteen-bit dense compute (TFLOPS): V100 125, A100 312, H100 989, B200 (HGX) 2,250, Rubin 4,000. Memory bandwidth (TB/s): 0.9, 2.039, 3.35, 7.7, 22 (Rubin preliminary datasheet, 22 June 2026; its product page gives 19.2). From the B200 on, one 'GPU' is two dies in one package, so the last two points compare two dies with one. Peak vendor figures, not measurements.

▼ Table view

Figure 5. NVIDIA flagship accelerators at fixed precision: sixteen-bit compute and memory bandwidth (V100 = 1)

Accelerator	16-bit compute	Bandwidth
V100	1x	1x
A100	2.5x	2.3x
H100	7.9x	3.7x
B200	18x	8.6x
Rubin	32x	24.4x

Figure 6. Growth from the V100 (2017) to Rubin (2026): the four-bit headline against sixteen-bit compute and memory bandwidth



V100: 125 TFLOPS FP16 tensor, 900 GB/s (NVIDIA V100 datasheet). Rubin: 35 PFLOPS dense NVFP4 (training figure), 4 PFLOPS dense BF16, 22 TB/s (NVIDIA Vera Rubin datasheet, preliminary, 22 June 2026). The sixteen-bit and bandwidth multiples are close; the 280-fold figure compares four-bit arithmetic with sixteen-bit. Peak vendor figures.

▼ Table view

Figure 6. Growth from the V100 (2017) to Rubin (2026): the four-bit headline against sixteen-bit compute and memory bandwidth

Quantity, V100 to Rubin	Multiple
Dense compute at the lowest precision each chip offers (FP16 to NVFP4)	280x
Dense compute at sixteen bits	32x
Memory bandwidth	24.4x

The dies explain the naming dispute. NVIDIA's datasheet describes each of Blackwell's two dies as "reticle-limited" — as large as the lithography used to print it allows — and NVIDIA's July 2026 technical blog says Rubin is likewise "constructed from reticle limited compute dies to achieve high density and efficiency", pairing them in one package: "These two dies are unified on a single package through a high-speed inter-die link." Once the thing sold as a GPU had become two pieces of silicon, whether to count it as one or two was a choice rather than a fact, and NVIDIA made the choice both ways.

The same blog is explicit about which resource Rubin's designers were chasing. It describes the chip's memory subsystem as providing "up to 22 TB/s of memory bandwidth: a 2.8x increase over Blackwell and Blackwell Ultra". That multiple rests on the 22-terabyte figure set against Blackwell's 8; on the 19.2 terabytes a second the product page has shown since September it is 2.4. The blog post was last modified on 6 August 2026, before the product page changed.

5. How the rest of the industry counts

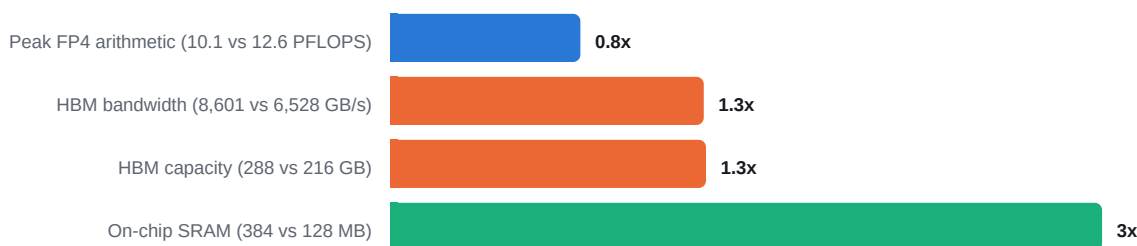
Every vendor chooses which of its numbers to lead with, and comparisons between vendors choose which of the rival's numbers to set against their own.

Competitors' comparisons inherit whichever NVIDIA figure they happened to read. AMD's page for its forthcoming Instinct MI455X states that the chip offers "50% more memory capacity and 6% more memory bandwidth vs. NVIDIA Vera Rubin GPUs"; its table sets AMD's 23.3 terabytes a second against 22.0 for Rubin, and its footnote dates the calculation to June 2026. Against the 19.2 on NVIDIA's product page since September the margin would be about 21%. The same page states that the MI455X offers "15% more OCP MXFP4 peak theoretical performance compared to NVIDIA Vera Rubin GPUs with FP4 datatype". The

accompanying chart sets AMD's 40 petaFLOPS against 35 for Rubin — NVIDIA's dense training figure, not the 50 petaFLOPS NVIDIA prints for inference. AMD's page does not say whether its own 40 is dense or sparse. NVIDIA's comparisons run the other way. The Rubin datasheet's chart of "10x More Tokens" against the previous HGX B200 system carries a footnote: "HGX Rubin NVL8 with Sparse NVFP4, HGX B200 with Dense NVFP4", and describes the result as "projected performance subject to change". A January 2026 NVIDIA technical blog labels the same 50-petaFLOPS Rubin figure "Transformer Engine compute" rather than "sparse", and gives the dense figure as 35; on a like-for-like dense basis, Rubin's four-bit arithmetic is 3.5 times Blackwell's rather than five. Google, for its part, sells Ironwood as one chip while its software presents it as two devices.

Google's newest designs fit the same pattern. On 22 April 2026 it described two eighth-generation TPUs with different jobs: TPU 8t, "optimized for massive-scale pre-training", and TPU 8i, for "sampling, serving, and reasoning". The serving chip has less peak arithmetic than the training chip — 10.1 petaFLOPS at four-bit precision against 12.6 — but more memory (288 gigabytes against 216), more memory bandwidth (8,601 gigabytes a second against 6,528) and three times the on-chip SRAM (384 megabytes against 128). Google explains the larger SRAM as room to "host a larger KV Cache entirely on silicon, significantly reducing the idle time of the cores during long-context decoding" — the attention cache, kept on the chip so the cores wait less. It gives no reason for the HBM figures, but the whole balance runs in the direction the roofline argument predicts for a chip built to serve answers.

Figure 7. Google's eighth-generation TPUs: the serving chip trades arithmetic for memory (TPU 8i as a multiple of TPU 8t)



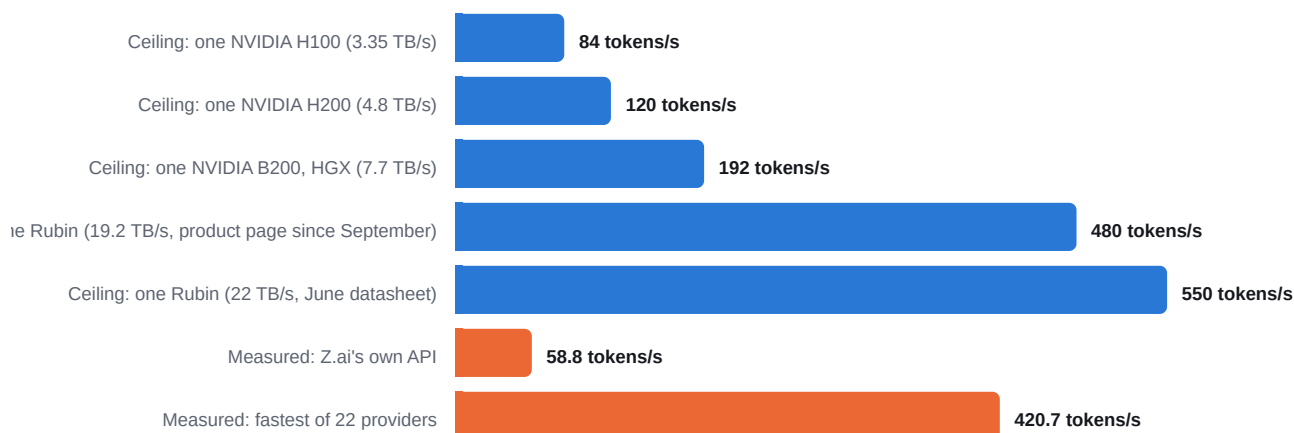
Source: Google Cloud Blog, 'TPU 8t and TPU 8i technical deep dive', 22 April 2026, 'at a glance' table, read 24 September 2026. TPU 8t is described as optimised for large-scale pre-training, TPU 8i for sampling, serving and reasoning. Vendor peak figures; neither chip has an independent published measurement.

▼ Table view

Figure 7. Google's eighth-generation TPUs: the serving chip trades arithmetic for memory (TPU 8i as a multiple of TPU 8t)

Specification	TPU 8i ÷ TPU 8t
Peak FP4 arithmetic (10.1 vs 12.6 PFLOPS)	0.8x
HBM bandwidth (8,601 vs 6,528 GB/s)	1.3x
HBM capacity (288 vs 216 GB)	1.3x
On-chip SRAM (384 vs 128 MB)	3x

The practical defence is to insist on two numbers together — dense operations per second at a named precision, and memory bandwidth in bytes per second — and then to look for a party measuring the quantity that matters. Neither is plentiful. Peak specifications are published only by the vendors themselves. Measurements that name the hardware, such as MLPerf's peer-reviewed results, which the vendors themselves submit, test standard benchmark workloads; measurements of a particular model as people actually use it seldom say what it runs on. Artificial Analysis, an evaluation firm that neither builds chips nor sells GLM-5.3, shows the model on its page, read on 24 September 2026, generating 58.8 tokens a second on Z.ai's own service; across the 22 providers it tracks, output speed ranged from about 43 tokens a second to 420.7, a spread of 874%. The providers' page does not name their hardware, although some label their precision ("Nebius (FP4)"). A provider exceeding the one-H100 bound of 84 tokens a second is not a contradiction of the arithmetic: the providers do not say what they run on, and it can mean a newer chip with more bandwidth, several chips working together, weights stored in fewer bits, or techniques that produce more than one token per pass. MLCommons, the industry consortium that runs the MLPerf benchmarks, published its Inference v6.1 results on 16 September 2026 with the statement that "NVIDIA Rubin and NVIDIA Vera Rubin NVL72 is in preview".

Figure 8. GLM-5.3, single-user generation: bandwidth ceilings on one chip against measured speeds

Ceilings are arithmetic, not measurements: memory bandwidth from each vendor document divided by about 40 GB, the approximate size of GLM-5.3's roughly 40 billion active parameters at eight bits (active count from Artificial Analysis, read 24 September 2026). They ignore that the full model (roughly 750 GB at eight bits per weight) needs several chips, the attention cache, batching, sustained-versus-peak bandwidth, and multi-token decoding. Measurements from Artificial Analysis's GLM-5.3 model and providers pages, read 24 September 2026; the providers do not disclose their hardware, and the fastest's configuration is unknown. The comparison is loose by construction.

▼ Table view

Figure 8. GLM-5.3, single-user generation: bandwidth ceilings on one chip against measured speeds

Chip ceiling (bandwidth ÷ 40 GB) or measurement	Tokens per second
Ceiling: one NVIDIA H100 (3.35 TB/s)	84 tokens/s
Ceiling: one NVIDIA H200 (4.8 TB/s)	120 tokens/s
Ceiling: one NVIDIA B200, HGX (7.7 TB/s)	192 tokens/s
Ceiling: one Rubin (19.2 TB/s, product page since September)	480 tokens/s
Ceiling: one Rubin (22 TB/s, June datasheet)	550 tokens/s
Measured: Z.ai's own API	58.8 tokens/s
Measured: fastest of 22 providers	420.7 tokens/s

6. What to watch

- **Which Rubin bandwidth NVIDIA stands behind once it ships.** The product page has given 19.2 terabytes a second per GPU since September; the June datasheet, the HGX page and the July technical blog still give 22. NVIDIA said in May that production shipments begin this autumn, and in August that racks were running at partners. Whether its documents converge, and on which number, and whether any independent party measures achieved bandwidth, will settle the figure on which every Rubin ceiling computed from it depends.
- **Rubin's exit from preview in MLPerf Inference.** MLCommons listed Rubin and the Vera Rubin NVL72 as "in preview" in its v6.1 results of 16 September 2026, and described v6.0 as "just six months ago"; if that cadence holds, the next round falls around the spring of 2027, a date MLCommons has not

published. Whether Rubin then appears as an available system is a check on NVIDIA's shipment claims published by a body other than NVIDIA, though the submission itself comes from the vendor.

- **Whether AMD re-bases its comparison.** AMD's MI400 page computes "6% more memory bandwidth" against Rubin at 22.0 terabytes a second, a calculation its footnote dates to June 2026. If the figure becomes about 21%, against NVIDIA's current 19.2, competitors will have started comparing on the revised memory number rather than the original one.

7. The idea to keep

The concept that makes chip announcements legible is arithmetic intensity — "operational intensity" in the roofline paper's own terms: how many operations a job performs for each byte it moves from memory. A modern accelerator is a very fast calculator at the end of a comparatively narrow road from a warehouse. The petaFLOPS on the datasheet describe the calculator; the terabytes a second describe the road; and whether a given job is limited by one or the other depends on its intensity. Generating text for a single user sits far to the memory-bound side, at about two operations per byte against ridge points in the hundreds, which is why memory bandwidth — the total the job can draw on — and not the headline petaFLOPS sets its pace.

Three questions follow for any hardware claim. How many operations a second, dense, at what precision? How many bytes a second from memory? And how many operations does the work in question perform per byte? The clearest statement of the framework remains its origin: Samuel Williams, Andrew Waterman and David Patterson, "Roofline: An Insightful Visual Performance Model for Multicore Architectures", *Communications of the ACM* 52(4), April 2009.