Forwards or Backwards

https://forwardorbackwards.com



Nvidia and 4nm - Blackwell, Rubin and the Node That Isn't

Watch on YouTube: https://www.youtube.com/watch?v=P99HCchkwKk

Chapters

1. A number the company never gave you
2. What Nvidia wrote in twenty twenty-two
3. The sentence from the foundry
4. Two spellings, and they are not the same thing
5. Twenty twenty-four, and the word gets louder
6. Twenty twenty-six, and the node disappears
7. What the four was actually doing
8. The limit that is actually there

9. Two dies pretending to be one
10. And then the units changed
11. The asterisk, and the ceiling
12. What to think when you see it

1. A number the company never gave you



Every account of the last four years of artificial intelligence hardware carries the same shorthand. Nvidia's chips are four nanometre chips. It appears in analysis, in procurement documents, in comparisons of one accelerator against another, and it is repeated with the confidence people reserve for specifications.

Here is what makes it worth an hour of your attention. Nvidia has never sold you a four nanometre chip, because the thing it bought from its foundry was never called that, and the company has been perfectly clear about this in public, in writing, on its own website, for four years.

This film is not an exposé. Nobody involved is hiding anything, and the interesting material is all published. What it is instead is a close reading — of one company's own architecture posts, across three generations of hardware, watching what they say about the process the silicon was made on.

In two thousand and twenty-two, Nvidia named the process. In two thousand and twenty-four, it named a custom one. And in two thousand and twenty-six, for its newest and largest part, it named nothing at all.

That progression is the first half of this film, and every step of it comes from Nvidia's own pages, read on the same day.

The second half is the part that matters more, because it answers the question the first half raises. If the node name is not the constraint, what is? And the answer turns out to be a physical limit almost

nobody outside the industry has heard of, which explains why the last two generations of Nvidia's biggest chips are not one chip at all.

And then, at the very end, the arithmetic. Because the performance figures changed units somewhere along the way, and reading them correctly is a skill.

2. What Nvidia wrote in twenty twenty-two



Start with Hopper, the generation that put the company at the centre of everything.

On the twenty-second of March two thousand and twenty-two, Nvidia published a long technical post about the architecture on its own developer site. It is still there. It contains, in plain language, the sentence that this entire subject depends on.

The chip, it says, is fabricated using the TSMC 4N process customized for NVIDIA.

Read it again slowly, because there are two claims in it and the coverage only ever carries one.

The first is the process name, 4N. The second is the phrase that follows it: customized for NVIDIA. This is not a standard product taken off a foundry's shelf. It is a variant, made to order, for one customer.

The same post gives the two figures that actually describe the chip. Eighty billion transistors. A die size of eight hundred and fourteen square millimetres.

Notice which of those three things is a measurement. The transistor count is a count of real objects. The die size is an area, in real units, of a real rectangle of silicon you could hold. And the node name is a product code.

Only two of the three are dimensions, and the one everybody quotes is not among them.

None of this required a leak, a teardown or an anonymous source. It is the first technical thing Nvidia published about the part, written by Nvidia, addressed to developers, still online four and a half years later.

The word to keep hold of is customized. Everything in the first half of this film follows from it.

3. The sentence from the foundry



So if 4N is a customised variant, a variant of what?

The answer is on the website of the company that actually operates the machines, and it is not a leak either. It is a press release.

On the twenty-sixth of October two thousand and twenty-one, Taiwan Semiconductor announced a process called N4P. Its own announcement describes it in a single clause that settles the whole question.

As the third major enhancement of TSMC's 5nm family, it says, N4P will deliver an eleven per cent performance boost over the original N5 technology.

Five nanometre family. Those are the foundry's words about its own four nanometre product, in its own press release, published openly, five years ago.

The rest of that release is worth having too, because it shows exactly the pattern the previous film on this subject found across the whole industry. Every figure is a ratio against a named predecessor. Eleven per cent more performance than N5. Six per cent more than N4. Twenty-two per cent better power efficiency than N5. Six per cent more transistor density than N5.

Four figures, two baselines, not one absolute dimension anywhere.

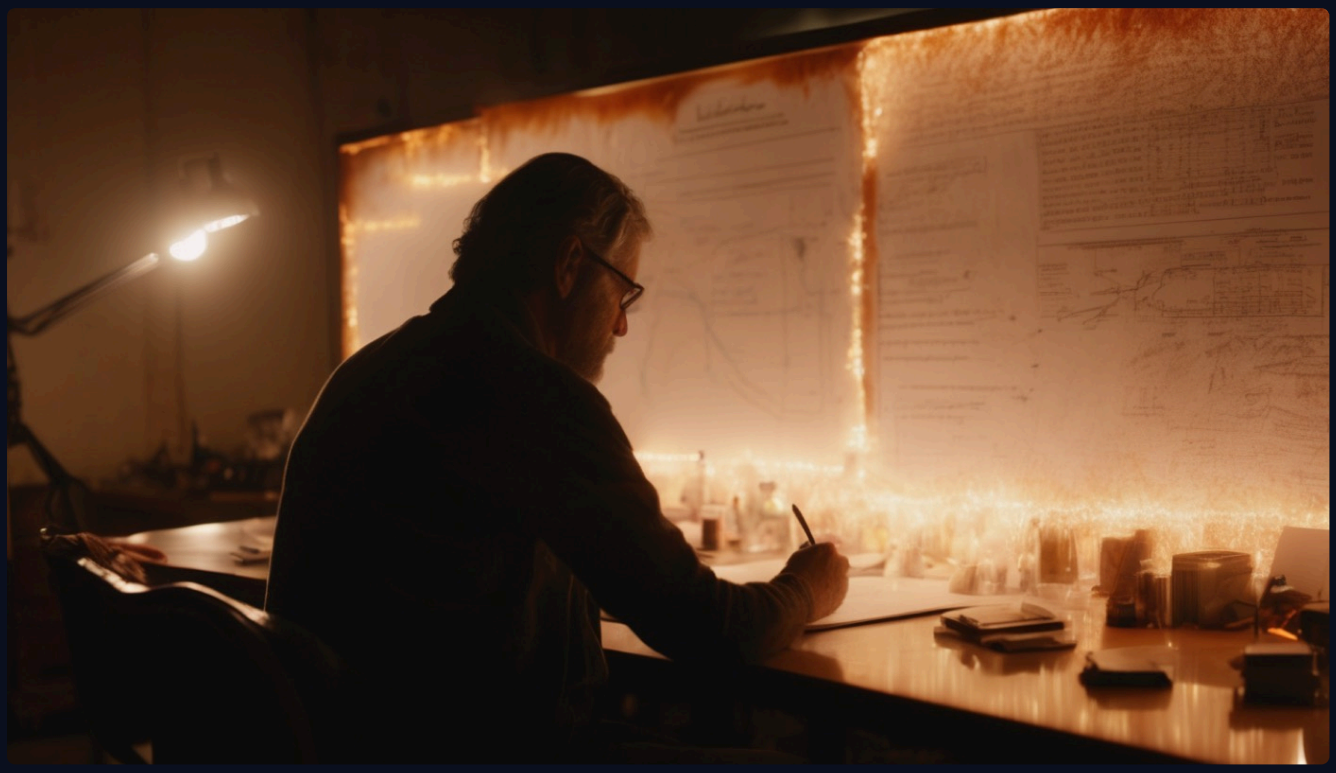
Now put the two documents side by side. The foundry says its four nanometre line is an enhancement of its five nanometre family. The customer says its chip uses a four nanometre process

customised for it.

So the number four, in the name of the process that made Hopper, is a refinement of a generation labelled five, tuned to order for one buyer. It was a commercial designation two steps removed from anything you could measure before Nvidia's marketing department ever saw it.

And the improvements underneath it are real. Twenty-two per cent better power efficiency is a substantial engineering achievement. It is simply not a length.

4. Two spellings, and they are not the same thing



There is a trap here that catches almost everyone who writes about this, and it is worth thirty seconds because getting it wrong makes the rest incoherent.

The foundry's products are called N4, N4P and N5. Letter first, number second.

Nvidia's process is called 4N. Number first, letter second.

Those are different strings and they name different things. N4 is a product in Taiwan Semiconductor's public catalogue, available to any customer who can pay for it. 4N is Nvidia's custom variant, and its specification has never been published by either company.

Write one when you mean the other and you have silently changed the claim from a bespoke process nobody outside two companies has seen, into a catalogue item with a public datasheet.

The pair get transposed constantly, including in otherwise careful reporting, which is how a great deal of confident commentary about Nvidia's manufacturing turns out to be commentary about a different process.

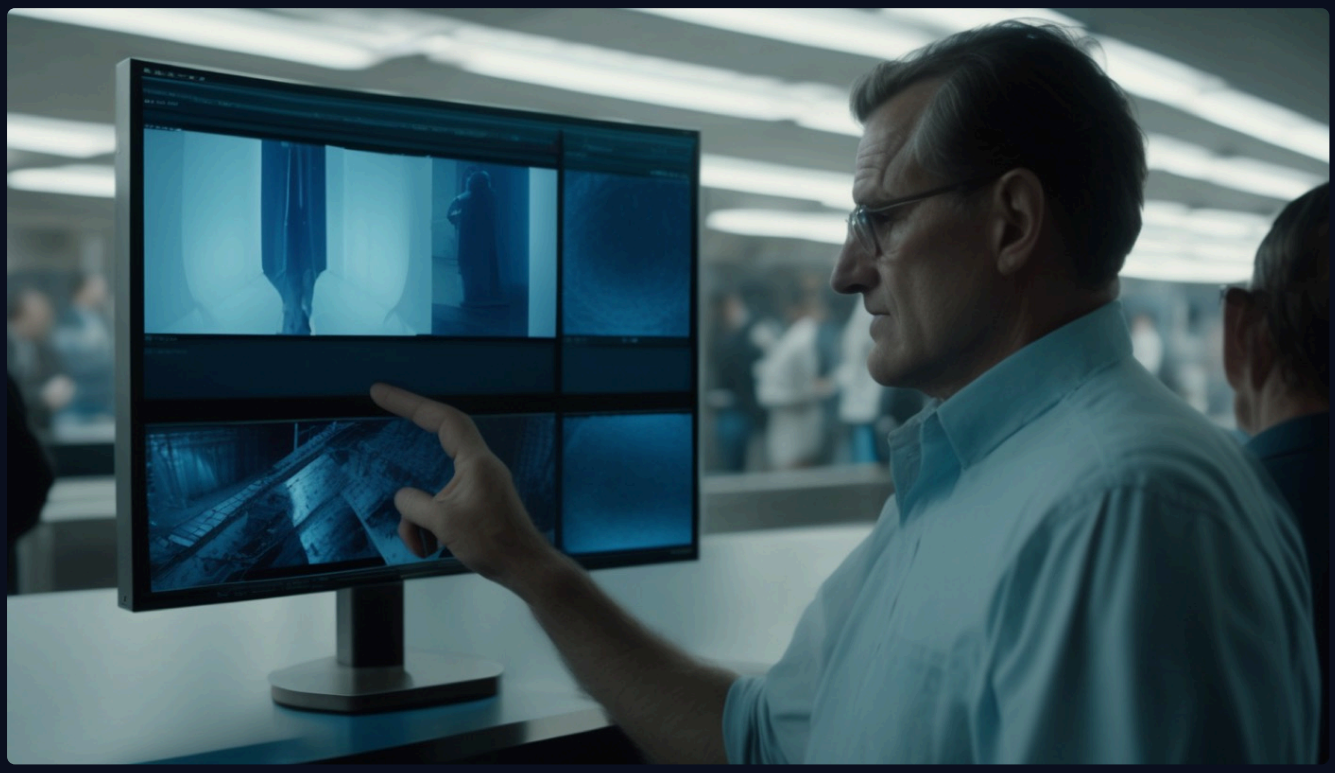
What follows from the distinction is the useful part. Because 4N is custom, there is no public specification for it. No gate pitch, no metal pitch, no cell height, no density figure. Not withheld — simply never published, because a bespoke process does not get a product page.

So when somebody compares Nvidia's four nanometres against another company's four nanometres, there is nothing underneath the comparison. Two product codes are being set against

each other as though they were two measurements, and neither one is a measurement, and one of them has no published parameters at all.

That is the state of the evidence, and it holds for every generation that follows.

5. Twenty twenty-four, and the word gets louder



Two years later, Blackwell.

Nvidia's own architecture page for it, read in September two thousand and twenty-six, keeps the same construction and strengthens it. The part is manufactured, it says, using a custom-built TSMC 4NP process.

Custom-built. Not customised this time — custom-built, and the process name has gained a letter.

Everything true of 4N is true here with more force. There is no public specification for 4NP as Nvidia uses it. The foundry's own N4P, the closest published relative, is the one described as the third major enhancement of a five nanometre family. Nvidia's version is a variant of that, and its parameters are not in circulation anywhere.

The same page gives the figures that are real. Two hundred and eight billion transistors. And then the phrase that the second half of this film is built on: two reticle-limited dies, connected by a chip-to-chip interconnect running at ten terabytes per second.

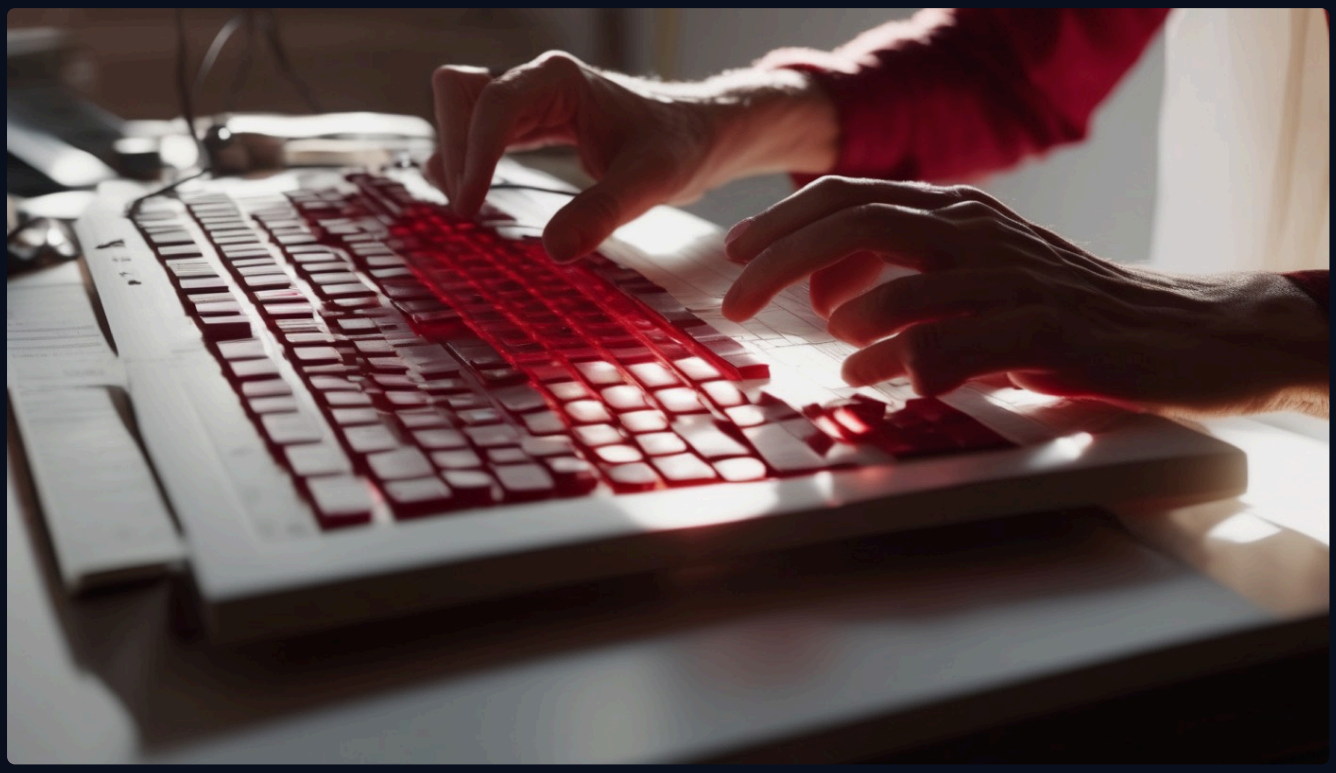
Hold that for a moment, because of what is missing from it.

Nvidia publishes a transistor count. Nvidia publishes an interconnect bandwidth. Nvidia publishes the number of dies and the fact that they are reticle-limited. What Nvidia does not publish, anywhere on that page, is a die area — and it published one for Hopper two years earlier.

So the disclosure is not simply thinning. It is changing shape. The company is telling you more about the architecture and less about the silicon, and the one physical dimension it did give you last time has quietly gone.

The node name is still there in twenty twenty-four. It is the last generation where it is.

6. Twenty twenty-six, and the node disappears



On the twenty-first of July two thousand and twenty-six, Nvidia published its architecture post for Rubin, the generation now in production. Three named authors, on the company's own developer site, in the same series as the Hopper post from four years earlier.

It is a detailed document. Three hundred and thirty-six billion transistors. Two dies, described as reticle limited compute dies, unified on a single package by a high-speed inter-die link. Up to two hundred and eighty-eight gigabytes of HBM4 memory, at up to twenty-two terabytes per second of peak bandwidth. NVLink 6 at three thousand six hundred gigabytes per second. A chip-to-CPU link at one thousand eight hundred. PCIe Gen 6 at up to two hundred and fifty-six.

That is a lot of specification. It is more architectural detail than the Hopper post carried.

And the process node is not in it.

That is a negative claim, so here is exactly how it was checked. The post was searched in September two thousand and twenty-six for the word TSMC, for N3, for three nanometre, for nanometre in any spelling, for process node, for foundry, and for fabricated. None of those terms appears anywhere in the document.

Not hedged. Not given as a range. Not attributed to a partner. Absent.

Trade publications report that Rubin is made on a three nanometre process, and they may well be right. But no primary source from either Nvidia or the foundry was found for it, and so it does not go

on screen in this film as a fact. An absence reported honestly is worth more than a number borrowed from somebody else's speculation.

Four years, three architecture posts, one company, one venue. The process named. The process named as custom-built. And then no process at all.

7. What the four was actually doing



So what was that number ever for?

The honest answer is that it was doing marketing work, and that this is not the same thing as lying.

A node name tells a buyer roughly where a process sits in a sequence — that it is newer than the last one, that it belongs to a particular tier of cost and capability, that a design targeted at it will land in a certain envelope of power and speed. All of that is real information. None of it is a length.

What the number does not tell you, and never did, is how large anything on the chip is. On a bespoke variant it does not even tell you which catalogue process it descends from, because the variant has no published parameters to compare.

And here is the part that makes the disappearance in twenty twenty-six make sense rather than look sinister. If the name never carried the information people were reading out of it, then dropping it costs the reader nothing. The transistor count is still published. The memory bandwidth is still published. The interconnect speeds are still published. Everything that can be checked is still there.

What went is the one item on the list that could not be checked by anybody.

You can read that as a company tidying up its own claims, or as a company retiring a label that had stopped doing useful work. Both readings fit the evidence, and this film is not going to pick one for you.

What neither reading supports is the idea that something is being concealed. A firm concealing its manufacturing does not publish three hundred and thirty-six billion transistors and the width of its inter-die link.

8. The limit that is actually there



If the node name is not the constraint these chips are built against, something is. And it is a limit with a name most people have never encountered, which is a shame, because it explains the shape of every large accelerator built in the last three years.

Recall how a pattern gets onto silicon. It lives on a mask, light is thrown at the mask, and the shadow is projected down through optics onto the wafer, shrinking on the way. That projected image has a maximum size. The machine can only throw so large a field before the optics stop being able to hold the pattern sharp across it.

That maximum printable field is called the reticle limit. It is not a manufacturing preference, a yield trade-off or a cost decision. It is the largest single rectangle of pattern the tool can print in one exposure, and no amount of money moves it.

Now look back at Hopper, and what Nvidia published about it in twenty twenty-two. A die of eight hundred and fourteen square millimetres. Trade reporting at the time put the theoretical maximum at eight hundred and fifty-eight, and that figure is journalism rather than a foundry specification, so treat it as approximate. But the direction is not in doubt, and Nvidia's own language for the generation that followed removes any need to trust a secondary figure.

Because when Nvidia described Blackwell, it did not say large dies. It said reticle-limited dies. That is the company's own word, for its own product, and it means the die is not as big as the designers wanted. It is as big as the printing equipment permits.

At that point the single-chip road ends. Not because the transistors could not be made smaller — the foundry has newer processes — but because the picture cannot be printed any wider.

Which raises the obvious question. Nvidia's transistor counts have gone up by a factor of four since Hopper. If the die cannot grow, where did they go?

9. Two dies pretending to be one



They went sideways.

Blackwell is two dies. Rubin is two dies. Neither is a chip in the sense the word usually implies, and both are sold, benchmarked and reasoned about as though they were.

Set the three generations out and the pattern is unmistakable. Hopper: eighty billion transistors, one die, eight hundred and fourteen square millimetres. Blackwell: two hundred and eight billion, two dies, each at the reticle limit. Rubin: three hundred and thirty-six billion, two dies, again reticle-limited.

The count roughly quadruples. The individual die does not grow at all, because it cannot.

What makes the arrangement work is the join. Nvidia's Blackwell page gives its speed: ten terabytes per second between the two dies. For Rubin the link has a name, NV-HBI, and Nvidia describes the two dies as unified on a single package.

Unified is doing real work in that sentence. The engineering goal of a link that fast is that software above it never has to know there are two pieces. A program is not written for a left die and a right die. It is written for one accelerator, and the interconnect has to be quick enough that the illusion holds under load.

That is the actual frontier, and it is nothing to do with nanometres. The hard problems are the width of that link, the substrate it crosses, the power delivered into two hot rectangles sitting millimetres

apart, and the cooling of the result.

Which is why the node name faded out of the disclosure just as multi-die packaging arrived. The interesting number stopped being which process printed the die. It became how many dies, how they are joined, and how fast the join runs.

10. And then the units changed



Now the second trap, and it is the one that costs people real money when they plan hardware.

At the Blackwell launch in March two thousand and twenty-four, the headline figure was twenty petaflops of artificial intelligence performance from a single part, against a maximum of four petaflops for Hopper before it. Five times, in one generation.

Both numbers are real. They are not measured in the same units.

The twenty petaflops figure is at FP4 — a four-bit floating point number format introduced with that generation. Hopper's four petaflops is at FP8, an eight-bit format. Halving the width of the number roughly doubles the throughput, because the hardware moves and multiplies half as much data per operation.

Trade coverage at the time did the like-for-like sum and published it, which is to its credit. Comparing the two parts at the same precision, at FP8, Blackwell offers around two and a half times the theoretical compute of Hopper.

So there are two ratios, and they are both honest. Around five times, if you allow the newer part to use its new narrower format. Around two and a half times, if you hold the format constant.

Neither is the wrong answer. They answer different questions. The first asks what the new hardware can do at its best; the second asks how much faster the same arithmetic runs.

What is not defensible is quoting the first ratio while describing it as the second — and that is the default in almost everything written about this hardware. A performance figure without its number format attached is not a specification. It is a headline with the units removed.

Rubin continues the convention, and the format has a name now: up to fifty petaflops of NVFP4 performance, in Nvidia's own words, on its own page.

11. The asterisk, and the ceiling



Two smaller habits worth learning to see, because both appear on Nvidia's own published pages and both change what a number means.

The first is sparsity. Go to Nvidia's specification table for the H100 today and the FP8 figure is three thousand nine hundred and fifty-eight teraflops. Directly underneath the table is a footnote, three words long: with sparsity.

Sparsity is a technique that skips arithmetic on values that are zero. Where a model has the right structure, the saving is real and the speed-up is real. Where it does not, the number above the footnote is not the number you get. The headline is a best case with its condition printed below it, honestly, in small type that almost never survives being quoted.

The second habit is the ceiling. Read Nvidia's own wording for Rubin and the phrase is up to fifty petaflops. Up to two hundred and eighty-eight gigabytes of memory. Up to twenty-two terabytes per second of bandwidth.

Up to is a ceiling, not a measurement. It describes the top configuration under favourable conditions. Strip those two words — as headlines routinely do — and a maximum silently becomes a typical figure.

None of this is deceptive on Nvidia's part. The footnote is printed. The words up to are written. The number format is named. Every qualification a careful reader needs is present in the source.

They are simply the first things lost when a figure is repeated, and they are lost every single time.

So the discipline is small and it is the same discipline in each case. Ask what precision. Ask whether sparsity is assumed. Ask whether the figure is a ceiling or a measurement. Ask what it is being compared against. Four questions, and the sources answer all four.

12. What to think when you see it



So, a node name on an Nvidia part. What should you make of it?

Not a dimension. Nothing on a Blackwell die is four nanometres, the company has never said there was, and for its current generation it does not name a process at all.

Not a comparison, either. Nvidia's 4N and 4NP are bespoke variants with no published parameters, and setting them against another company's similarly numbered process compares two product codes rather than two measurements.

What it is, at most, is a rough position in a foundry's sequence — and the foundry itself files its four nanometre line under a five nanometre family, in a press release anyone can read.

Meanwhile the numbers that do hold still are the ones that get quoted least. Eighty billion transistors, then two hundred and eight, then three hundred and thirty-six. Eight hundred and fourteen square millimetres of silicon, and then a limit the printing equipment will not cross, so that the last two generations are each two dies and a link running at ten terabytes per second between them.

Those are counts, areas and rates. Every one is published by the company, every one has a unit, and not one of them is a marketing label.

The pattern is the same one the previous film on this subject found in the industry at large, arriving here in a sharper form. The label got vaguer — from a named process, to a custom-built one, to nothing — while the published engineering got more specific, more granular and easier to check.

Nvidia stopped telling you the number that never meant anything. It is still telling you all the ones that do.

Sources and credits

Primary sources

- Nvidia, 'NVIDIA Hopper Architecture In-Depth', developer.nvidia.com, published 22 March 2022 – 'The full GH100 GPU that powers the H100 GPU is fabricated using the TSMC 4N process customized for NVIDIA'; 80 billion transistors; 'a die size of 814 mm2'. PRIMARY, and the only generation for which Nvidia publishes a die area.
- Nvidia, 'Inside NVIDIA Rubin GPU Architecture: Powering the Era of Agentic AI', developer.nvidia.com, 21 July 2026 (Eduardo Alvarez, Vishal Mehta, Farshad Ghodsian) – 336 billion transistors; 'reticle limited compute dies', two dies unified on one package over NV-HBI; up to 288 GB HBM4 at up to 22 TB/s; NVLink 6 at 3,600 GB/s; NVLink-C2C at 1,800 GB/s; PCIe Gen 6 up to 256 GB/s; 'up to 50 petaflops of NVFP4 performance'; '10x more agentic throughput per unit of energy than NVIDIA Blackwell'. PRIMARY. THE PROCESS NODE IS ABSENT – searched 18 September 2026 for TSMC, N3, 3nm, nanometer/nanometre, process node, foundry and fabricated; none appears.
- TSMC, 'TSMC Expands Advanced Technology Leadership with N4P Process', pr.tsmc.com/english/news/2874, published 26 October 2021 – verbatim: 'As the third major enhancement of TSMC's 5nm family, N4P will deliver an 11% performance boost over the original N5 technology'; +6% over N4; +22% power efficiency and +6% transistor density against N5. PRIMARY, and the single cleanest fact in the film.
- Tom's Hardware, Jarred Walton, 'Nvidia's next-gen AI GPU is 4X faster than Hopper', 18 March 2024 'Blackwell B200 gets to that figure via a new FP4 number format, with twice the throughput as Hopper H100's FP8 format'; 'if we were comparing apples to apples and sticking with FP8, B200 only offers 2.5X more theoretical FP8 compute than H100'; H100 'has a die size of 814 mm2, where the theoretical maximum is 858 mm2'.
- Tom's Hardware, Aaron Klotz, 'Nvidia CEO confirms Vera Rubin NVL72 is now in production', 6 January 2026 – SECONDARY. Vera Rubin announced in full production at the CES 2026 keynote; 50 PFLOPS NVFP4 inference and 35 PFLOPS NVFP4 training; transistor count 1.6x Blackwell; 288 GB HBM4; availability second half of 2026. Used only for the production milestone; the reconciliation 336/208 = 1.615 is recorded in the fact sheet as a cross-check, not shown as a figure.

Not regulated financial advice.