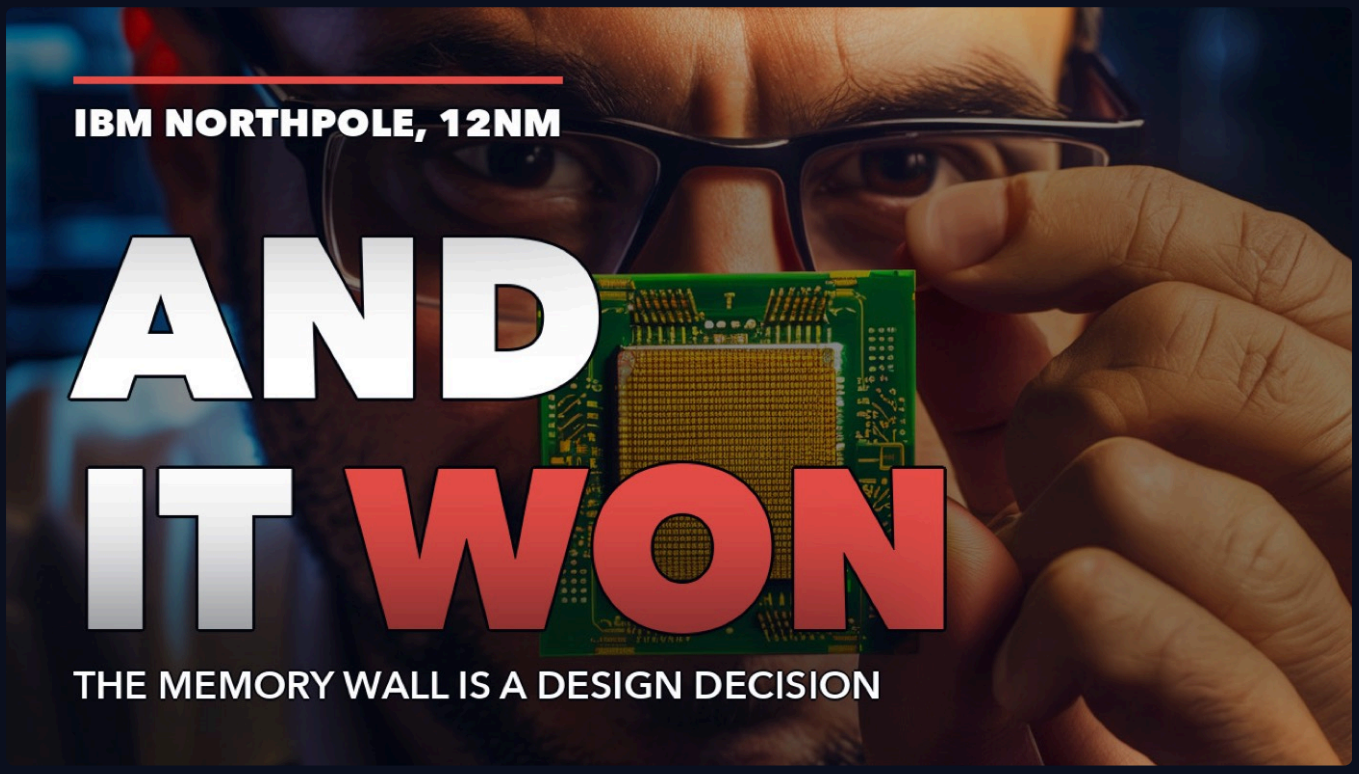
Forwards or Backwards

https://forwardorbackwards.com



IBM NorthPole — Inference on a 12nm Chip

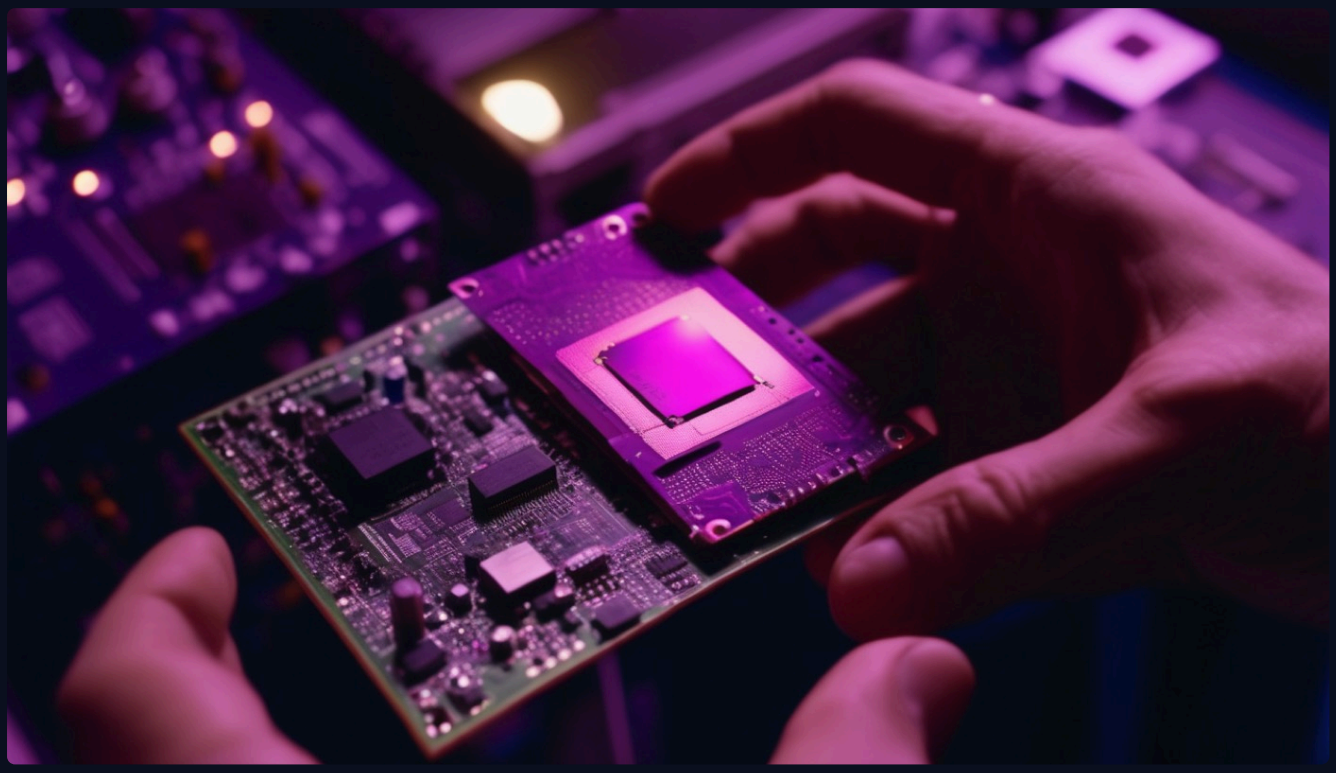
Watch on YouTube: https://www.youtube.com/watch?v=0UIKl3z-yUg

Chapters

1. The chip that went the wrong way
2. The problem is not arithmetic
3. What NorthPole took out
4. The chip itself
5. Precision, and why it is not cheating
6. Twenty twenty-four: one server
7. The two numbers this film will not repeat carelessly
8. Twenty twenty-five: the rack
9. Two point eight is not slower than one
10. What is missing from the evidence

11. The constraint nobody has removed
12. What this actually demonstrates
13. Forwards or backwards

1. The chip that went the wrong way



Almost everything you have heard about making chips faster is a story about making them smaller. Five nanometres, then four, then three. Each step costs more than the last, and each step is announced as though the alternative were standing still.

This film is about a chip that went the other way, and won.

NorthPole is an IBM research processor built for one job: running a neural network that has already been trained. Not training it — running it. It was described in the journal Science on the nineteenth of October, 2023, by Dharmendra Modha and his colleagues at IBM Research.

It is built on a twelve nanometre process, at GlobalFoundries. In 2023, when it was published, twelve nanometres was not a leading-edge node. It was not close to one. It was the sort of process you would use for a part where cost mattered more than speed.

And on the problems it was designed for, it outran chips built on far smaller nodes.

The interesting thing is not that this happened. It is why. NorthPole is not faster because its transistors switch faster — they do not. It is faster because of something it does not have. The design removes a component that every other processor of its kind is organised around, and almost everything else about the chip is a consequence of that single absence.

So this is a film about a decision rather than a device. Someone looked at the thing everybody treats as a fixed cost of computing, and asked what the machine would look like if you simply refused to

pay it.

2. The problem is not arithmetic



To see what NorthPole removed, you have to be honest about what a processor running a neural network actually spends its time doing.

The arithmetic is not hard. Running a trained network is mostly multiplication and addition, arranged in a fixed order, with no surprises in it. A modern processor can do an enormous number of those every second — far more than it usually gets the chance to.

Because the numbers are not where it needs them.

The weights of a trained network are stored in memory chips outside the processor. To do a calculation, the processor has to fetch the numbers, wait for them to arrive, do the arithmetic — which takes almost no time at all — and then fetch the next batch.

The fetching is slower than the arithmetic. Not slightly. By a wide margin, and the margin has been widening for decades, because arithmetic has got faster much more quickly than memory access has.

This is the memory wall, and it is the central fact of the field. The consequence is that a great deal of the engineering in a modern accelerator is not about calculating. It is about hiding the wait: caches, prefetching, deep pipelines, running many things at once so that something useful happens while the rest of the machine is waiting for numbers.

All of that is real engineering and it works. But look at what it is. It is an elaborate structure built to conceal a delay, rather than to remove it.

Which raises the obvious question, and it is the question NorthPole is an answer to. What if the numbers never had to travel?

3. What NorthPole took out



NorthPole runs inference without off-chip memory.

That is the whole idea, and it is worth sitting with, because it sounds smaller than it is. There is no bank of memory chips beside the processor holding the network's weights. During inference, the chip does not go off the die to fetch anything.

Instead the memory is distributed across the chip itself, placed beside the compute that uses it. Two hundred and fifty six cores, each with its own memory next to it, wired to its neighbours.

The network is not loaded from somewhere when it is needed. It is already there, spread across the chip, sitting where it will be used.

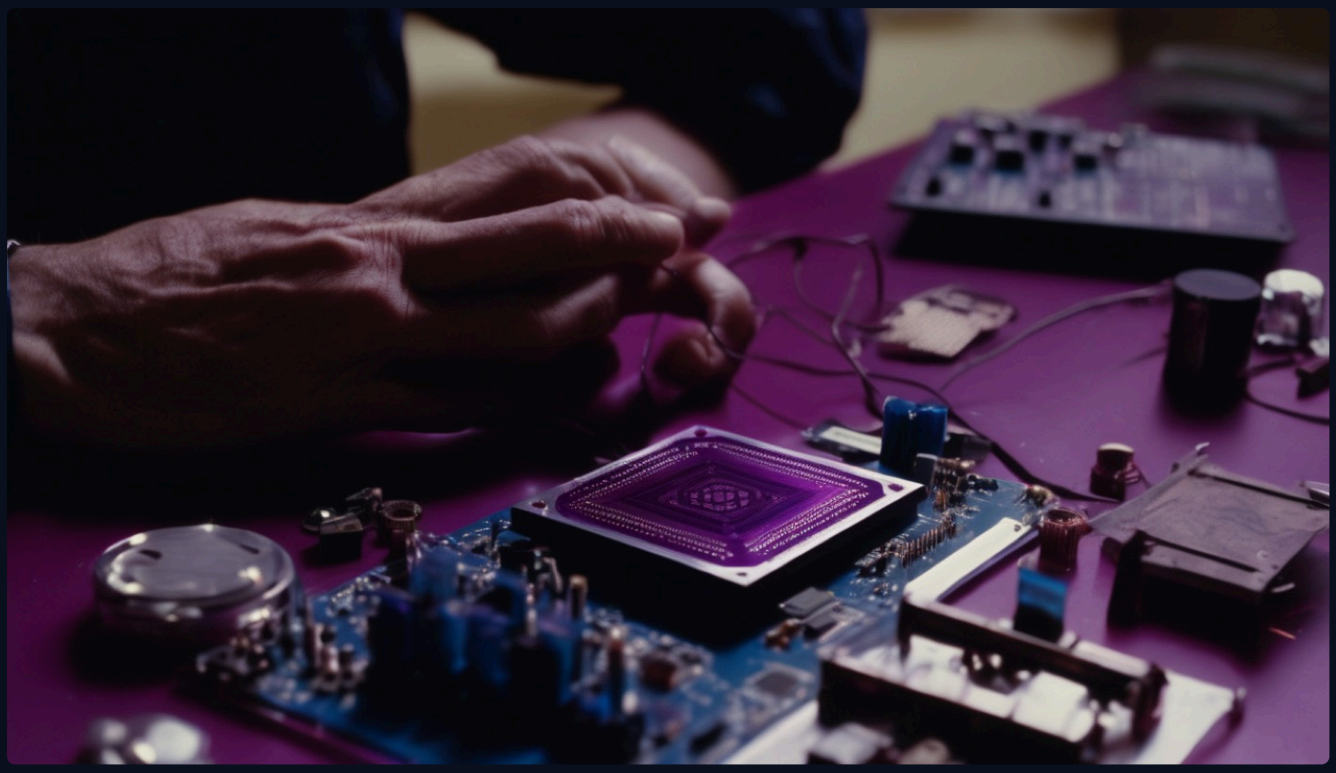
And once you make that choice, the memory wall does not get hidden. It stops existing — for this job, under these conditions. There is no fetch in the inner loop, so there is no wait to conceal, so the entire apparatus built for concealing it becomes unnecessary and can be spent on something else.

This is why the process node stops being the interesting number. Shrinking transistors makes arithmetic faster, and arithmetic was never the bottleneck. NorthPole did not make the fast part faster. It deleted the slow part.

Modha has a phrase for the result, and it is a fair summary of the claim: architecture trumps Moore's law.

There is a cost, and the film will come back to it, because it is the real limitation of the approach. If the weights have to fit on the chip, then the chip's memory sets the size of what you can run. That constraint shapes everything that follows.

4. The chip itself



So, the specifications, each with its source, because this is a film that intends to be checkable.

From the Science paper of October 2023, and the architecture paper presented at the International Solid-State Circuits Conference in 2024:

Twenty two billion transistors, in an area of eight hundred square millimetres, on the GlobalFoundries twelve nanometre process. That is a large die. Not an exotic one — a large one, on a mature process, which is a combination that tends to be affordable in a way that leading-edge parts are not.

Two hundred and fifty six cores.

Two hundred and twenty four megabytes of memory on the chip. That figure divides into a hundred and ninety two megabytes across the core array, and thirty two megabytes of frame buffer.

Two hundred and twenty four megabytes sounds small, and against the memory attached to a modern accelerator it is small. That is the constraint from the last chapter, stated as a number. It is also the entire point: this memory is not smaller memory, it is closer memory, and closeness is what was bought.

Two thousand and forty eight threads. Over four thousand wires crossing each core in each direction, which is what it takes to move data between cores when you have decided it will not move anywhere else.

And on-chip memory bandwidth of thirteen terabytes per second, per chip. That qualifier matters and the film will keep repeating it: per chip.

5. Precision, and why it is not cheating



There is one more design decision, and it is the one most likely to sound like a trick.

NorthPole does not work in the high-precision numbers that training uses. It supports eight bits, four bits and two bits, mixed as needed.

Each core does two thousand and forty eight operations per cycle at eight-bit precision. At four bits, four thousand and ninety six. At two bits, eight thousand one hundred and ninety two. Halve the precision, double the throughput.

Now, reducing precision is not free, and anyone who tells you otherwise is selling something. A four-bit number holds less than an eight-bit one, and a network run at low precision can give different answers from the same network run at full precision.

But there is a real finding underneath this, and it is well established rather than controversial: a trained network is far more tolerant of low precision when it is being run than when it is being trained. Training needs fine gradations because it is making small adjustments. Running does not.

So the precision is chosen for the job, not shaved to flatter a benchmark. And it has a second effect that matters more than the arithmetic speed. Lower precision means the weights take fewer bits, which means more of the network fits in those two hundred and twenty four megabytes — which is the constraint that governs everything on this chip.

That is worth naming clearly, because it is the honest shape of the design. The precision choice and the no-off-chip-memory choice are not two clever ideas. They are one idea. Low precision is what makes keeping everything on the chip possible at all.

6. Twenty twenty-four: one server



Specifications are a claim about what a chip could do. The question is what it does.

On the twenty sixth of September, 2024, IBM Research published results for NorthPole running a large language model. Sixteen NorthPole cards, in a single two-rack-unit server.

The model was a three-billion-parameter variant of the open-source Granite-8b-code-base-4k model, run with four-bit weights and activations.

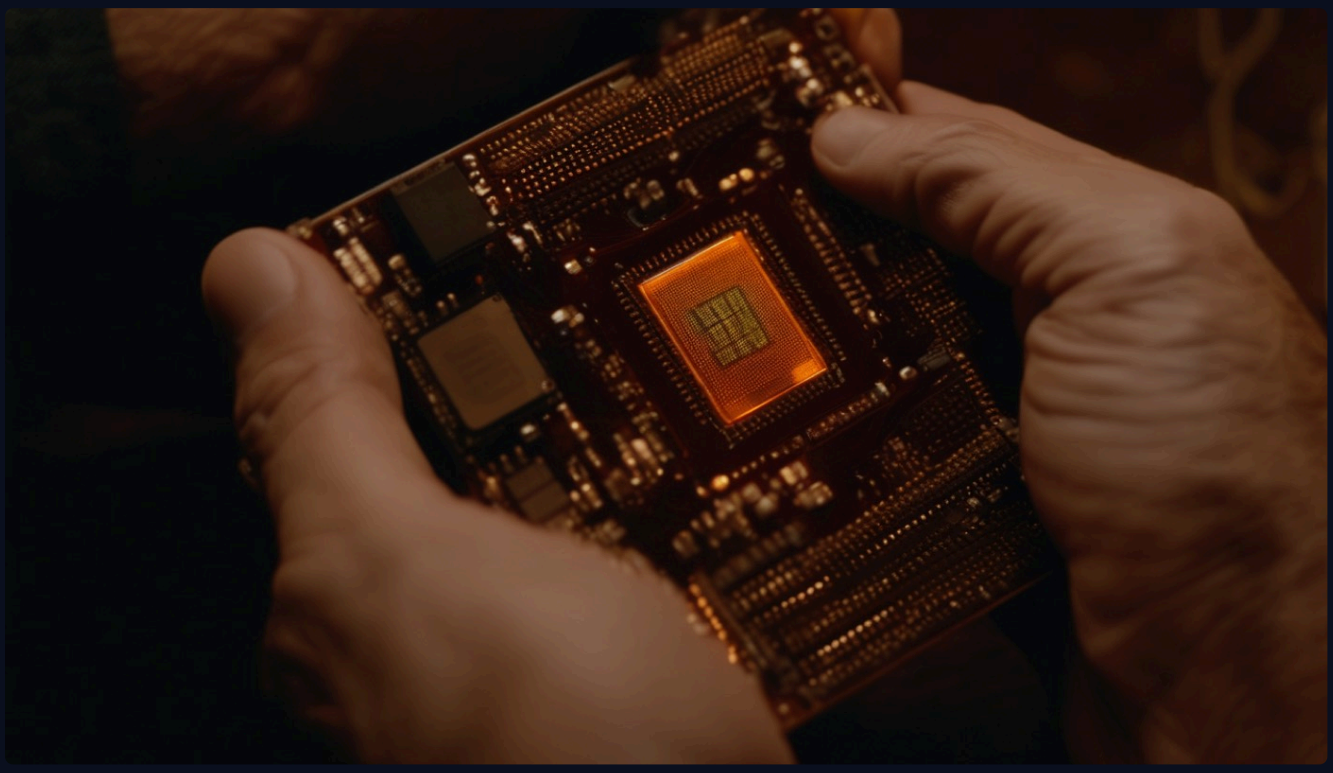
That description is deliberately long, and the film will always say it the long way, because there is a trap in the short version. Called "a three-billion-parameter Granite model", it sounds like one thing. Called "Granite-8B-Code-Base", it sounds like another, larger thing. They are the same model: a three-billion-parameter variant derived from the eight-billion-parameter code base model. Shortening it either way misleads, so it does not get shortened.

The results, on that configuration: token latency below one millisecond, and a system throughput of twenty eight thousand three hundred and fifty six tokens per second.

IBM reports that this came with close to the accuracy of the full-precision model — which is the claim the four-bit decision has to survive, and the reason the quantisation is named in the same breath as the result.

Modha's summary of what he takes the result to mean: what is essential here is qualitative orders of magnitude in improvement.

7. The two numbers this film will not repeat carelessly



The 2024 results came with two comparisons that are the most quotable numbers in this entire story, and they need handling.

Forty six point nine times faster than the next most energy-efficient GPU. Seventy two point seven times more energy efficient than the next lowest-latency GPU.

Those are extraordinary figures. Two things about them are also true, and they go on screen together with the numbers rather than in a footnote.

The first: these are IBM's measurements of IBM's chip, published by IBM. That does not make them wrong. Vendor benchmarks are often accurate. It does mean they are not independent, and nothing in this film's research found an independent benchmark of NorthPole. Every performance figure you have heard so far traces back to IBM or to IBM's authors.

The second, and it is the sharper problem: the comparison GPUs are not named. The phrases are "the next most energy-efficient GPU" and "the next lowest-latency GPU", which tells you the shape of the comparison but not who is in it. A multiplier against an unnamed part is not a checkable claim, and the difference between a well-chosen comparison and a badly-chosen one is the difference between a finding and an advertisement.

So: the numbers are reported here, attributed, with both caveats attached. What they support is that IBM measured a large advantage in a comparison IBM constructed. What they do not support is the

bare sentence "NorthPole is forty seven times faster than a GPU", and this film does not say that sentence.

8. Twenty twenty-five: the rack



The next result is larger, and its provenance is different in a way worth stating.

On the twentieth of November, 2025, thirty one authors at IBM submitted a paper to arXiv describing a scaled NorthPole system. This is a preprint — it has not been through peer review, and the film says so here, the first time it uses it.

The system: two hundred and eighty eight NorthPole cards, across eighteen two-rack-unit servers.

One hundred and fifteen peta-ops, at four-bit integer precision. Three point seven petabytes per second of memory bandwidth across the system.

Thirty kilowatts of power. Seven hundred and thirty kilograms. A footprint of zero point six seven square metres — a single forty-two-unit rack.

Thirty kilowatts in two thirds of a square metre is the figure to hold onto. That is the claim that the approach survives contact with a real data centre: not a laboratory curiosity on a bench, but a rack with a power draw and a floor area that a facilities engineer could plan around.

And what it runs: three simultaneous instances of IBM's Granite-3.3-8b-instruct — eight billion parameters — at two thousand and forty eight tokens of context, serving twenty eight simultaneous users, at a per-user inter-token latency of two point eight milliseconds.

The same rack can be reconfigured: eighteen instances of a three-billion-parameter model, or a single instance of a seventy-billion-parameter model. The constraint from chapter three has not gone

away. It has become a dial.

9. Two point eight is not slower than one



Put the two results next to each other and something looks wrong.

2024: below one millisecond per token. 2025: two point eight milliseconds per token.

The newer system is eighteen times the size of the older one, and by that reading it got almost three times slower. If you spotted that, you were right to stop — and the answer is that the two numbers are not measuring the same thing, and should never have been put side by side in the first place.

The 2025 figure is a per-user inter-token latency with twenty eight users being served simultaneously, on a two hundred and eighty eight card rack, running an eight-billion-parameter instruct model.

The 2024 figure is from a sixteen card server, a different year, a different and smaller model, and a different number of users.

Serving twenty eight people at once and serving one person are different problems, and the latency you can promise each of twenty eight people is not comparable to the latency you can achieve for one. A number without its configuration is not a slow number or a fast number. It is not a number about anything.

This is why every latency figure in this film appears on screen with its user count and its card count in the same frame, and why you have not seen the two of them in one shot. It would have been the easiest graphic in the film to make, and it would have been false.

10. What is missing from the evidence



A film that has just spent a chapter on careful reading owes you the same treatment of its own sources. So here is what this story does not have.

It does not have an independent benchmark. Every performance figure here – the 2023 chip results, the 2024 server, the 2025 rack – comes from IBM or from IBM's researchers. No third party was found in this research who has run NorthPole and published what they measured.

It does not have named comparisons. The most quoted numbers in the whole story are measured against GPUs nobody has identified.

And there is a third absence, which is the most interesting of the three, because it is IBM's own.

The 2025 paper is the newest, largest and most formally presented result in this story: thirty one authors, a full system, submitted for review. It contains no GPU comparison at all.

You can read that two ways, and the honest thing is to say both. It may mean simply that a paper describing a system's architecture had no reason to benchmark against somebody else's product. Or it may mean that the comparison got harder between 2024 and 2025, in a period when the parts on the other side of it were improving quickly.

What it does not mean is that the multipliers have been withdrawn or disproved. Nothing here says that. It means the strongest and most recent evidence for NorthPole is an absolute claim – this much

compute, this much bandwidth, this many users, in this much space, at this power — and not a ratio against a rival.

That is, if anything, the sturdier kind of claim. It is just not the kind that makes a headline.

11. The constraint nobody has removed



Return to the thing set aside in chapter three, because it is the limit of the whole approach.

If the weights live on the chip, the chip's memory decides what you can run. Two hundred and twenty four megabytes per chip is what NorthPole has, and no amount of architectural cleverness makes that number larger.

The answer, visibly, is more chips. Sixteen cards for a three-billion-parameter variant in 2024. Two hundred and eighty eight cards to hold a seventy-billion-parameter model as a single instance in 2025.

So the memory wall was not abolished. It was moved. Instead of a wall between the processor and its memory chips, there is now a question of how many cards you need before the model fits, and how they talk to each other when it does not fit on one.

That is a genuinely better place for the problem to be — moving data between cards in a rack you designed is a more tractable engineering problem than waiting on DRAM in an inner loop. But it is a relocation, not a disappearance, and a design that runs a seventy-billion-parameter model by spreading it over two hundred and eighty eight cards has a scaling story it will eventually have to answer for.

Models have not been getting smaller.

12. What this actually demonstrates



Strip out the multipliers and the rack specifications, and there is a claim left that does not depend on any of the contested figures.

A chip on a mature, cheap, twelve nanometre process, from a foundry nobody thinks of as leading-edge, ran inference competitively against parts built on far more advanced nodes. That happened. It was published in Science, presented at ISSCC, and nobody has disputed it.

The reason it happened is that the industry's dominant metric – how small are the transistors – was measuring the wrong thing for this workload. The bottleneck was never the speed of the arithmetic. It was the distance to the numbers. A process shrink improves the first and does almost nothing for the second.

That is the finding, and it is worth more than the benchmarks. Not "NorthPole beat a GPU", which is a claim about a product. But: on this workload, where the memory is mattered more than how small the transistors were. Design choice beat process node.

That generalises well beyond one research chip from one laboratory, and it is the reason this story is interesting even if every multiplier in it turns out to have been generously chosen.

13. Forwards or backwards



So which was it.

By the measure the industry has used for fifty years, NorthPole is a step backwards. Twelve nanometres in an era of single-digit nodes, a large die, mature process, nothing exotic about the manufacturing at all.

By the measure that actually governs the job, it is a step forwards, and a substantial one — because it identified a cost everyone else had accepted as permanent and declined to pay it.

Both of those are true simultaneously, and that is the useful thing here rather than a paradox to be resolved. The direction of progress depends entirely on which axis you are measuring, and the industry had settled on an axis. NorthPole is what it looks like when somebody checks whether the agreed axis is still the one that matters.

It has not displaced the GPU. There is no evidence presented in this film that it has, and it would be dishonest to imply it. It is a research prototype, benchmarked by the people who built it, against competitors they did not name.

But it demonstrated something that does not need a benchmark to be true: that the memory wall is a design decision rather than a law of nature, and that a chip which refuses it can be built, and can be scaled into a rack that draws thirty kilowatts and fits in two thirds of a square metre.

Going backwards on the one number everybody watches turned out to be the way forwards on the one that mattered.

That is worth knowing whoever eventually wins.

Sources and credits

Primary sources

- Modha et al., 'Neural inference at the frontier of energy, space, and time', Science, published 19 October 2023, doi 10.1126/science.adh1174 - the chip: 22 billion transistors, 800 mm2, 12 nm GlobalFoundries, 256 cores, 224 MB on-chip (192 MB core array + 32 MB frame buffer), 2,048 threads, 2,048/4,096/8,192 ops per core per cycle at 8/4/2-bit, over 4,096 wires crossing each core in each direction.
- '11.4 IBM NorthPole: An Architecture for Neural Network Inference with a 12nm Chip', ISSCC 2024, IEEE - the architecture paper.
- IBM Research blog, 'NorthPole LLM inference results', published 26 September 2024 - 16 cards in one 2U server; a three-billion-parameter variant of Granite-8b-code-base-4k at 4-bit weights and activations; latency below 1 ms per token; 28,356 tokens/s system throughput; 46.9x and 72.7x against UNNAMED GPUs. Quotes Dharmendra Modha: 'What is essential here is qualitative orders of magnitude in improvement.'
- arXiv 2511.15950, 'A Scalable NorthPole System with End-to-End Vertical Integration for Low-Latency and Energy-Efficient LLM Inference', 31 authors, IBM, submitted 20 November 2025 288 cards across 18 2U servers; 115 peta-ops at 4-bit integer; 3.7 PB/s; 30 kW; 730 kg; 0.67 m2 42U rack; 3 instances of Granite-3.3-8b-instruct at 2,048 context serving 28 simultaneous users at 2.8 ms per-user inter-token latency; or 18 instances of a 3B model, or one 70B instance. This paper is also the primary source that settled the die area, the per-chip bandwidth unit and the identity of the 2024 model.

Not regulated financial advice.