

What GLM-5.3 Flash running on Chinese hardware actually means

August 29, 2026 · [Martin Alderson](#)

Z.AI [confirmed](#) that their most recent model release was running all inference on Chinese manufactured hardware. While no doubt an impressive feat, Western companies still have a huge advantage that I can't see changing quickly.

Where is Chinese AI hardware at?

To start with, it's worth looking into *where* Chinese AI hardware is. I'm focusing entirely on the HiSilicon parts - the most competitive parts from Huawei. There are (many, actually) other manufacturers building AI hardware, but it's widely believed that they are no further ahead than HiSilicon, so I think that for brevity it's a fair starting point.

One caveat before I go further: Z.AI didn't actually name a chipmaker, and didn't publish throughput or power numbers either. Nobody has independently verified the claim. So I'm assuming HiSilicon here because it's the only plausible candidate at that scale, not because anyone has confirmed it.

It's also worth mentioning that the US export restrictions ([CSIS has a good overview](#)) of high end AI hardware have made this an enormous priority, understandably, for the Chinese. And it's definitely worth mentioning that finding accurate sources for many of the numbers I'll cite are difficult to be confident in, so take the exact numbers with a pinch of salt.

The current 'scale-up' series of HiSilicon chip, the 910c series, pairs 96GB of HBM 2e memory with two compute dies, probably achieving something like 1.6PFLOP/s of INT8 compute with ~3TB/sec of memory bandwidth, at around 600W.

In essence, this is substantially behind even the H100 from Nvidia, which is now 4 years old. These are [around 60%](#) as fast as the H100, and has various other footguns (no native FP8

support for example), which probably restrict efficiency further for many use cases.

The next generation 950-series doesn't meaningfully increase compute as far as I can see, but does use domestically produced HiZQ/HiBL HBM memory. Interestingly the cards are configured in two variants - the 950PR and 950DT, with the former focusing on prefill and the latter on decode. In reality, the two products are very similar, but the prefill variant using slower HiBL memory vs the decode HiZQ memory. It does however support more quantisation types, like FP8.

The constraints

I think this shows the limitations of what Chinese hardware can do - at least for the near future.

Yes, they can run inference, but so can *many* sets of hardware now - AMD, Google and Amazon all have competitive solutions, and OpenAI are making significant progress on their [Jalapeño inference chip](#), which in the [first published benchmarks](#) did 1.5-1.9x the work per watt of Nvidia's GB300. Inference hardware while no doubt complex, is a pretty *solved* problem right now with a lot of competition - and that's before you bring in the Cerebras and Groq approach chips.

The wall that these Chinese hardware manufacturers are hitting is the lack of viable EUV (extreme ultraviolet) fabrication. This is the next generation silicon manufacturing process from ASML and it is *extremely hard*. I'd really, *really* recommend reading [Chip War](#) by Chris Miller for the full story, but regardless until there is significant progress on this - and by significant progress, I don't mean the [reverse engineered prototype](#) in a Shenzhen lab. I mean reliable, *scale* production.

The industry would be astonished if they got this to scale production before 2030. Bear in mind the Shenzhen prototype hasn't produced a working chip yet, and the more optimistic forecasts have them doing that around 2030 - volume production is a further step beyond it. It took ASML 25 years to figure out this technology - and a good 5+ years of this was scaling it up from the lab to "real" production lines. While China no doubt has incredible engineering talent *and* the ability to reverse engineer some of ASML's work, it's still a daunting challenge.

Without EUV it is not possible to go (much) below the "7nm" fabrication size. Without being able to go below that size, you quickly hit a wall in thermal efficiency, and you reach a point where you simply cannot make the chip(s) any bigger or faster because you cannot expel the heat quickly enough.

Added to that, the additional export restrictions on HBM memory to China are clearly causing significant issues, hence the strange use of two different home grown memory technologies in the 950-series - no doubt because they can't produce enough fast (which is still comparatively *slow*) memory.

These are really the same base constraint - without EUV manufacturing technology you can't produce the latest generations of very fast HBM memory either.

But maybe this doesn't matter?

Clearly the approach China is taking is instead of really looking for solid incremental leaps in compute and memory from better manufacturing techniques, the idea is to build *a lot* of them. Even if your fastest chips are at best 5 years behind the latest Nvidia GPUs, you can just build 10 times as many for the same overall inference capacity. And it really is roughly 10x - not against the H100 I was comparing to above, but against what Nvidia actually ships today. A Rubin VR200 is somewhere around 35PFLOP/s of dense FP4 with 22TB/sec of HBM4 bandwidth. The 910c is 60% of a four year old H100; Rubin is another order of magnitude past that.

No doubt China is uniquely positioned in being able to do this - with *enormous* power generation capacity to power this, and huge quantities of skilled engineering and manufacturing labour to build the facilities and cooling required.

But really, it's far from ideal. As models get larger, you have to split them over more and more underpowered sets of hardware. Another problem is it makes the models *slow* - Z.ai's own API is noticeably slower than Western providers serving the same weights.

The bit I keep coming back to though is power. And here you have to be careful, because 10x the throughput gap is *not* 10x the power bill - the 910c pulls about 600W against something like 2000W for a Rubin part. Divide the spec sheets and you get a much less dramatic 2-3x on both compute per watt and bandwidth per watt.

But the spec sheets flatter the 910c. 96GB a chip, against the 288GB or more you get on current Western parts, leaves much less room for KV cache, which forces smaller batches, and decode throughput per watt falls away badly at small batch sizes. Add a less mature software stack, and the interconnect and cooling overhead of running 10x the chips, and 5x worse on tokens per watt feels about right to me. If anything that's the charitable end.

Which matters because electricity is usually reckoned to be [10-20% of the total cost](#) of running a GPU cluster, with hardware amortisation dominating. Multiply that by five and power goes from a small component of costs to something like half your total bill. That's fine when you have China's generation capacity and you're happy to treat the difference as a strategic subsidy. It's a lot less fine if you ever want to sell inference into a competitive global market on price.

Small models getting better doesn't rescue this either. They help, obviously - a 30B model serving a task that used to need a 300B one is a real saving. But it's a saving both sides get - that smaller 30B model still runs 10x as fast on Western hardware, so the ratio between Chinese and Western hardware efficiency stays exactly where it was.

And *assuming* China doesn't have some huge breakthrough in fabrication technology - which as I said before is highly unlikely - it's probable that the gap between Western and Chinese AI hardware will widen if anything.

So, to round up - yes it's an impressive feat that they've managed to do this, but there are some hard constraints on efficiency that are unlikely to be solved any time soon. And yes, China *could* overcome it by sheer quantity, but it's a subpar solution that has real impact on the speed, capacity and economics of their inference.

RELATED

[Are we dismissing AI spend before the 6x lands?](#)

[What's going on with Gemini?](#)

[Why on-device agentic AI can't keep up](#)

[← Back to Home](#)

