SUPRAGRAPHOS

Compositional Cognition, Structural Intel

RESEARCH BRIEF

07 SEPTEMBER 2026 / SEMICONDUCTORS & AI

Huawei's τ Scaling Law

A Structural Read, and an Independent Check

From silicon geometry to power constraints, packaging leverage and diverging toolchains.



7 September 2026

About this brief

This document is structured into two parts.

Part One is our read of Huawei's own material: the τ scaling theory paper and the follow-up thermal paper published 4 September, plus the press coverage around the Kirin 2026 launch.

Part Two tests that read against two outside documents we did not consult* while forming it: a technical rebuttal from Chinese Academy of Sciences researchers, and an investment-strategy note from Invesco.

Our substrate is face value. We take every reported measurement as reported. We are not equipped to verify silicon and we do not try. What we do is read structure: what an argument is built on, what it needs but never supplies, whether its evidence supports the claim it is attached to, and where its parts fail to meet.

Our hedging rule follows from that, and we apply it without exception. We state directly what a document contains, omits, claims, or fails to reconcile, because that is verifiable by reading. We attribute everything about physics and manufacturing to whoever claimed it. Huawei reports. The Academy researchers derive. Invesco argues. Where we reason beyond the documents, we say so.

One limit on what Part Two proves. Both outside documents were published before Huawei's 4 September paper, and both predate our reading of them. So this is independent derivation followed by corroboration, not prediction.

*Our **methodology** can be found towards the end of the report.

CONTEXT NOTE / A

Seven terms, before we start

Nothing below requires an engineering background, but seven words do a lot of work in this story.

Node. The industry names manufacturing generations with a number: 7nm, 5nm, 2nm. The number stopped literally measuring anything years ago. Treat it as a generation label. Smaller means newer, more expensive, and made on more advanced machinery.

EUV lithography. The machines that print the finest circuit patterns onto silicon. One company in the world makes them, in the Netherlands, and they cannot be sold to China. This single fact is why the rest of this document exists.

Die. One rectangle of silicon: the chip itself. A phone's main chip carries several specialized blocks on the same die, each doing a different job. The four that matter here are the **CPU** (general-purpose work), the **GPU** (graphics), the **NPU** (AI processing), and the **DSP** (signal processing).

Transistor density. Transistors per square millimeter. The industry's headline progress number for sixty years.

Power versus power density. Total watts consumed, versus watts per square millimeter. These come apart, and the difference is the hinge of this entire story. A chip does not fail because it drew too many watts. It fails because one small area got too hot. You can lower a chip's total power while making one corner of it hotter, and that corner is what breaks.

Iso-performance. Comparing two chips doing exactly the same amount of work, so the only variable left is what it cost them. When Huawei says "41% less power," they mean at iso-performance: same job, less energy.

Tier, or layer. A sheet of working circuitry. A conventional chip has one. Huawei's technique uses two, stacked and bonded face to face. Their roadmap goes to more.

Part One: The Structural Read

Bottom line

Huawei has a real engineering result sitting inside a much larger theory, and the two deserve separate verdicts.

The engineering is about shortening the distance data has to travel inside a chip, and they now report measurements supporting it on the exact component critics expected to fail.

The theory claims to be a successor to Moore's Law as the industry's organizing principle, and it is not finished enough to be one.

The most durable idea in the whole body of work is a geometry argument that would still be true if the theory were discarded tomorrow.

What Huawei Claims, and What It Means Downstream

For sixty years the industry sold one thing under one name: smaller transistors.

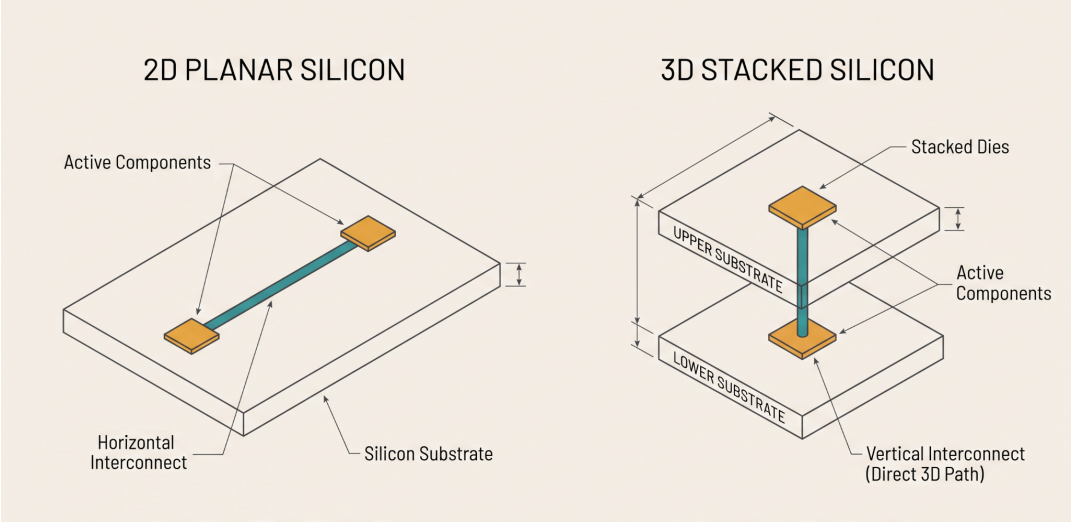
Huawei's reframing is that nobody ever actually wanted smaller. They wanted sooner. Smallness was simply how sooner has been bought so far, and shrinking has stopped paying off.

For Huawei it stopped abruptly, because export controls cut the company off from the most advanced manufacturing equipment. Their response is to change the unit of measurement: count time, not nanometers. Then anything that saves time counts as progress, and shrinking becomes one option rather than the only one.

The technique is to stop building flat.

A circuit that used to sprawl across a single plane is folded into two stacked layers, bonded together closely enough that the designer treats the pair as one continuous slab rather than as two chips talking to each other.

FIG. 01 / PLANAR VERSUS STACKED SILICON



Same circuit, same manufacturing process, same total transistors. The only thing that changed is how far the signals have to go.

SUPPLIED ARTWORK / EDITORIAL CROP

The obvious objection is heat.

Stack working circuits on top of working circuits inside a sealed handheld device and the outcome seems inevitable: more heat in less space, temperature climbs, the chip throttles or cooks. That objection is what the September paper exists to answer, and their argument is that it assumes the wrong power model.

Most of a chip's energy is not spent computing. It is spent moving bits along wires. A wire behaves like a capacitor smeared out over its length, so the further a signal has to travel, the more energy it costs to send, and at modern manufacturing scales the wires, not the transistors, dominate the electricity bill.

Their own analogy, which is a good one: for most office workers the largest energy expenditure of the day is not the job, it is the commute. Desk work is comparatively cheap; transport dominates. Fold the circuit and you move everyone's home closer to the office.

A second effect compounds the gains.

Folding frees up room for more transistors within the same footprint, and those spare transistors can be spent running more copies of a circuit side by side. Work split across many copies can be done more slowly by each of them, which means each can run at a lower voltage, and energy use falls with the square of voltage. So the first effect saves in proportion; the second saves in proportion squared.

Huawei reports that the second is where most of the benefit came from.

REPORTED MEASUREMENTS / ISO-PERFORMANCE

~2/3	~3/4	+~50%
less NPU power	less NPU power density	transistor density
Huawei-reported	Huawei-reported	Huawei-reported

APPROXIMATIONS FROM SOURCE TEXT / NOT THIRD-PARTY MEASUREMENTS

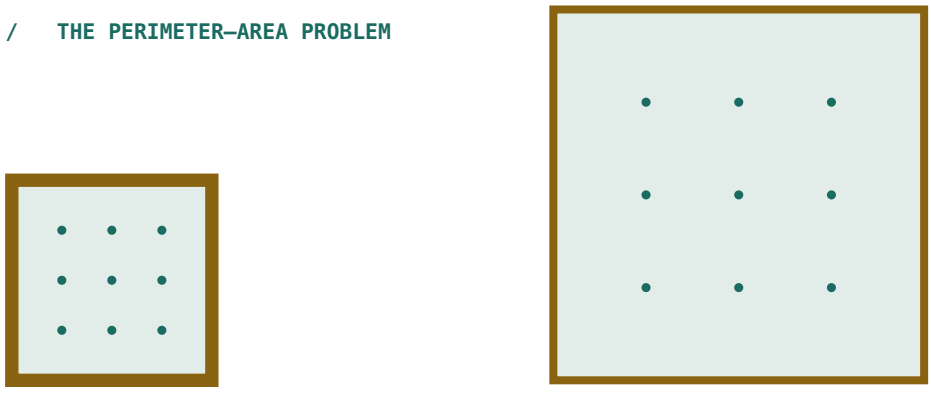
The reported results are consistent with that story and, notably, are strongest exactly where the objection was loudest. On the NPU, the AI block, which is both the densest and among the hottest parts of the chip and the one skeptics most expected to melt, Huawei reports roughly two-thirds less power and roughly three-quarters less power per unit of area at matched output. Transistor density across the chip rose by about half in a single generation on an unchanged manufacturing process, a gain they say previously took three years of shrinking to achieve.

What that means for anyone who is not a chip designer: when an AI assistant streams a reply to you word by word, the chip is mostly not thinking. It is waiting for data to arrive from memory. That is why buying more raw computing power does not make an assistant feel faster. Hardware built this way would make the same model start answering sooner, stream faster, cost less per answer, and stop penalizing long conversations so heavily.

It would not make the model smarter. Nothing here improves a model's capability. It makes the same capability cheaper and quicker, which is a different thing and, commercially, a more immediately bankable one.

What Holds Up

FIG. 02 / THE PERIMETER–AREA PROBLEM



SMALLER DIE LARGER DIE

Computing grows with area; off-chip access remains constrained by the edge.

CONCEPTUAL GEOMETRY / NOT TO SCALE

The shape argument. A building's capacity is set by its floor space, but everything and everyone enters through doors on the outside wall. Add floors and you have not added doors, and past a certain size the doors, not the floor space, decide how many people the building can actually serve.

A chip works the same way. Its computing power grows with its area, while memory, data connections and electrical current all enter through its edge. Area outruns edge, permanently, and no improvement to individual transistors touches this, because it is a problem of shape rather than of components. The only real fix is to stop relying on the edge and start using the top and bottom faces.

Our analysis frameworks separate load-bearing claims from decorative ones by testing whether each would survive the removal of a document's central proposition, one we derive structurally rather than accept from the author's own framing. This claim survives cleanly: it neither depends on the new measure nor supports it. It would be equally true in a world where τ scaling had never been proposed.

The bonding threshold. Stacking chips is not new. TSMC, Intel and Samsung have all been refining it for years. What is new in this work is a specific claim about when stacking changes character, and it turns entirely on how finely the two layers are stitched together.

CONTEXT NOTE / B

Why the stitching is the whole story

Return to the building. Two floors, and the question is how people move between them.

One staircase. You must assign whole departments to a floor, because every trip between floors is expensive and there is one route. Sales upstairs, engineering downstairs. Nobody's desk can straddle the two.

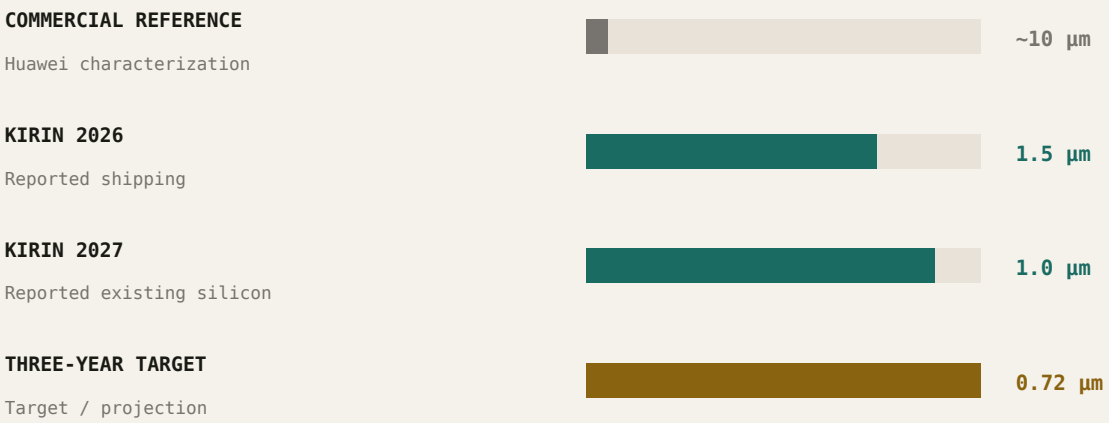
A staircase at every desk. Now it does not matter which floor anyone sits on. You place each person wherever the work is best served, and the building behaves like one large open floor that happens to be folded in half.

That is the difference Huawei is claiming. When the vertical connections between the two silicon layers are sparse, the design problem is "which whole functional blocks go on which layer," and it must be decided by hand, block by block. When the connections become nearly as fine as the ordinary wiring on top of a single chip, the problem becomes "place any individual logic element anywhere in the two-layer volume," which is a different and much larger design space.

The numbers give the scale of the claim. Huawei reports that advanced commercial 3D chips today place their vertical connections roughly 10 micrometers apart, which works out to fewer than a million links per 100 square millimeters. They report Kirin 2026 shipping at 1.5 micrometers, which is about 50 million vertical connections, of which 10 to 15 percent carry signals. They say Kirin 2027 silicon already exists at 1 micrometer and more than 100 million connections, and that within three years they expect to reach 720 nanometers, matching the pitch of the chip's own top wiring layer, at more than 200 million connections.

Taken at face value, that is roughly fifty times more connections per unit area than they describe as current commercial practice, with a further fourfold claimed ahead.

FIG. 03 / VERTICAL BOND PITCH



A finer pitch permits a denser set of short inter-layer pathways.

LOG-SCALED VISUAL GUIDANCE / VALUES AS STATED IN THE BRIEF

This is the most concrete and most checkable claim in the entire body of work, and in our reading it is the actual payload. Everything else is scaffolding around it.

What is Oversold

The theory has no arithmetic. The paper states that total system time is some function of the time at each layer, and never says what that function is. Not a sum, not a product, not a bounded combination. Just "a function of."

It is a company announcing that it will henceforth measure everything in customer satisfaction without saying how to add the sales team's number to the support team's. As a shared vocabulary it works and is useful, putting a physicist and a system architect in the same meeting talking about the same quantity. As something an engineer can calculate with and sign off against, it does not yet exist.

Moore's predecessor in this role, Dennard scaling, was a closed set of relationships you could actually compute.

The demonstrations do not validate the metric. Our frameworks check whether two lines of argument actually meet where a document implies they do, and here they do not. The stacking result, the shape argument and the networking improvements would all be exactly as true, and exactly as valuable, under the old way of measuring progress. The material reads as though its engineering wins confirm its new yardstick. On our reading they confirm the geometry, and the yardstick is still waiting.

And the theory cannot price energy. This deserves room, because it is the most consequential thing in the document and the authors themselves now agree.

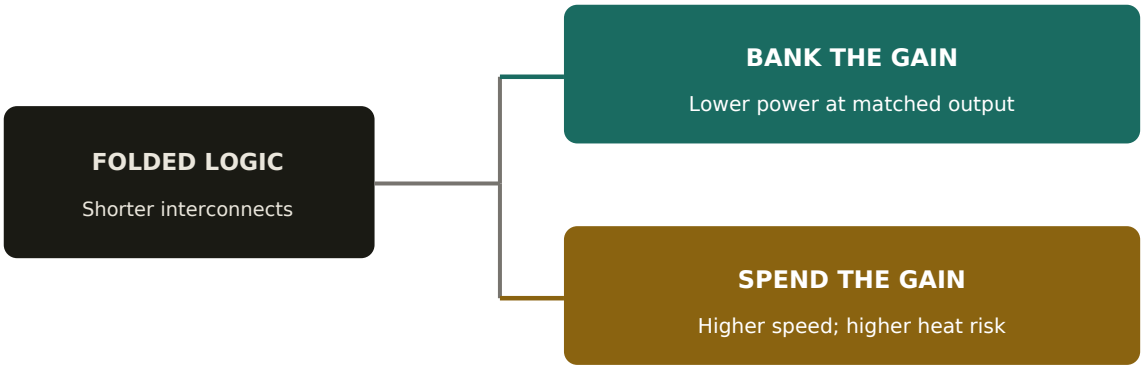
τ scaling tells you to make things faster. It does not tell you what to do with the speed you gain. That gap is not a detail. It is the whole commercial question.

When folding shortens the wires, the chip can do the same work in less time. The designer then faces a choice.

1. **Bank the gain**: keep performance where it was, drop the voltage and the clock, and the chip runs cooler and cheaper. Or:
2. **Spend the gain**: hold the voltage where it was and let the chip run much faster and much hotter.

Both are permitted. Nothing in the theory prefers either.

FIG. 04 / ONE GEOMETRY, TWO OPERATING CHOICES



The reported cooler outcome is a choice within the design space.

CONCEPTUAL STRUCTURE / NOT A PERFORMANCE PLOT

Huawei banked it. In their words, τ is a time scaling law and not an energy law, and a folded system running faster while burning more power breaks none of its rules and would still drain your battery by lunchtime. They call spending recovered speed on lower power a "companion discipline," and say it is a choice they plan to make each generation.

We would echo their emphasis and add one observation.

A discipline is a decision made once per product, by a company weighing its own incentives. It is not a property of the law, and it does not transfer with the technique. Their own figures show both settings on the same silicon: run flat out, the folded AI block delivers 141% more throughput than its predecessor, and its power per square millimeter rises above the chip it replaced. On the CPU the same maneuver pushes power density more than 40% above the planar baseline.

The cool chip is a choice Huawei made. The hot chip was equally available.

Their projections outrun their measurements. The July material put the frequency roadmap at one level; two months later it is a full step higher with no new mechanism given. The measured content in the September paper is disciplined and specific; the forward-looking content is not.

The gap between the two registers is worth a raised eyebrow.

The discipline is right for phones. It is unlikely to survive a data center. This is our own reading rather than anything the papers claim, and we offer it as inference.

Huawei reports that Kirin 2026 runs cooler than its predecessor. We take that as reported, and the benefit to a phone owner is real and easy to state: the phone does the same work while drawing less from the battery, so it lasts longer between charges. In a handheld device, battery life is what enforces the choice to bank the gains. Nobody ships a phone that wins benchmarks and dies by lunchtime.

A data center has no equivalent enforcement. There, output is revenue, and the same dial gets turned the other way. What happens when it is turned is not speculation on our part; it is in the tables. Power density goes above the chip it replaced, by roughly 7% on the AI block and by more than 40% on the CPU. The technique still works. The cool result does not follow it.

There is a second reason to expect the data-center case to differ.

Part of the phone result came from the folded blocks physically shrinking, which left surrounding room for heat to spread sideways before it has to travel up and out. A data-center processor is already built at the largest size that can be manufactured in one piece, so it has no such slack, and commercial pressure there would spend any recovered area on more computing rather than on thermal margin.

Our frameworks flag cases where a demonstration and the claim it supports sit at different scales of size, power and time. This is one. The demonstration is a handheld part drawing a few watts. The claim reaches to gigawatt facilities. And the incentive that produced the favorable result does not travel with the technique.

What It Means for the AI Build-Out

Value is moving toward memory and packaging, and this is already visible in what things cost. Trade estimates put high-bandwidth memory at close to half the manufacturing cost of a current top-end AI accelerator, with memory and advanced packaging together making up the clear majority of the parts cost. The expensive part of an AI chip is increasingly not the computing.

Huawei's material predicts that the deliberate separation of processor from memory, an arrangement dating to the 1980s that created an entire industry, is now reversing under pressure.

Here is our reading of what that adds up to.

The τ proposal is a currency substitution. A path that was not merely expensive but unavailable to Huawei when progress is priced in nanometers becomes available when it is priced in time. And the unit in which an industry prices progress is also the unit in which its choke points are drawn.

Two consequences follow, and they are the point.

1. The first is **demand**. If time is the thing being bought, then the things that buy time become where the money goes: wafer bonding, advanced packaging, putting memory physically closer to logic, faster interconnect. The cost shift already visible in an accelerator's parts list is the leading edge of that, and it accelerates with every additional player who adopts the same move. Nothing here requires Huawei to be right about the theory. It only requires enough of the industry to conclude that the remaining gains are in integration rather than in miniaturization, which is a conclusion several of them have already reached independently.
2. The second is **leverage**. Export controls are effective against lithography because a single company in one country makes the machines. Bonding equipment, packaging capacity and memory supply are concentrated too, but differently and less completely. Re-denominating progress into a unit whose inputs sit in different hands does not defeat the controls, but it does change the shape of what they can reach. That is a strategic fact independent of whether the underlying physics claim holds.

Power, not speed, is the ceiling. In the United States, generation and storage capacity waiting in grid connection queues now runs to roughly two thousand gigawatts, more than the country's entire installed capacity, with typical waits measured in years, while data-center demand is projected to nearly double by 2030.

The binding constraint on AI right now is getting electricity to the building.

This is a theory about latency arriving in an era whose headache is joules. Our frameworks surface what an argument requires but never supplies, and this one requires an energy account. By its authors' own admission, it does not contain one.

And there is a more subtle reason it may not happen quickly.

The material is candid that the approach needs design software that does not currently exist, and calls building it the single most important investment of the coming decade. The September paper goes further, making thermal control explicitly conditional on tools that can interleave hot and cool circuitry across layers.

But the organizations best placed to build that software are the ones who would be disadvantaged if the approach were widely adopted. Whoever switches sides does not merely have to fund new infrastructure. They have to open a moat they spent decades digging, and invite everyone through it. Their incumbency is denominated in the old unit; re-denominating is, for them, a write-down.

This is a weiji moment in the exact sense: the Chinese word pairs danger with the pivot point at which a situation turns. What makes it unusual is who is standing at the pivot. Huawei, the party with the least to lose from a change of measure, is the one proposing it. The parties with the most to lose are the ones being asked to build the instrument that would make it real.

Where Our Initial Read was Wrong

Across our analysis of Huawei's papers, several of our early positions changed, especially once the September paper appeared. Those are not errors so much as new evidence arriving. There is one finding of ours, though, that was simply wrong when we went back to check it.

We flagged an arithmetic inconsistency that turned out to be ours, not theirs. We noted that a large density gain on an unchanged manufacturing process could not be squared with the density formula the paper supplies.

It squares immediately once you notice what the "per square millimeter" is being measured against. Density is transistors divided by area, and the area Huawei divides by is the footprint of the folded stack, seen from above. Fold the circuit into two layers and there are now two sheets of silicon sitting under each square millimeter of footprint, so density per square millimeter rises without a single transistor having to shrink. We read the figure as a claim about a flat chip, which it was not, and had to correct it later.

We believe any professional with sufficient expertise in the matter would not have stumbled over the same pitfall. That is worth stating plainly: it is evidence for not merely a human in the loop, but an expert in the loop.

How to Read Huawei's Numbers

We take Huawei's numbers at face value because we do not have the expertise to validate or dispute them. But every performance figure in this document is Huawei's own, measured internally, on a product they are shipping this month. Do not silence the voice at the back of your head saying these are vendor claims. They should be held as such until someone outside the company measures a shipping part.

One internal inconsistency survives our analysis: the networking latency improvement is stated three different ways across the material, spanning a range of roughly fivefold between the smallest and largest versions of the same result. As a general pattern, the block-level measured data is considerably more consistent than the system-level and forward-looking claims.

There is also a marker in the other direction, and it is the strongest signal of good faith in the whole body of work. They volunteer the case that went against them.

One processing block, the DSP, saw its power per square millimeter move the wrong way in the first generation, rising by about a quarter, because its footprint shrank faster than its power consumption fell. They disclose it, explain exactly why, and cite later silicon that corrects it. Documents written purely to position do not usually surface their own counterexamples.

Part Two: The External Corroboration

Two outside documents, neither consulted while Part One was formed. Both were published before Huawei's 4 September thermal paper, which makes them a useful record of informed opinion at the moment the heat question was still open, and which also limits how much the first of them can be asked to settle.

What each document is

Sun and Wang, "From Moore's Law to the Volume-Inverse Law." Posted 24 August on preprints.org, explicitly not peer reviewed. The authors are at institutions of the Chinese Academy of Sciences, and one of them reports two decades in transistor process and device development at Intel. The paper is analytical rather than empirical: its own data statement says no datasets were generated or analyzed. So this is a derivation from stated assumptions, offered by people with directly relevant industrial experience.

Invesco, "An alternative model: China's low-cost AI strategy." July 2026, a marketing communication for professional investors, explicitly labeled as not financial advice, from a house arguing that Chinese technology equities are attractively valued relative to global peers. It has a directional interest in the thesis it advances, and its figures predate the September data.

The technical counter-thesis

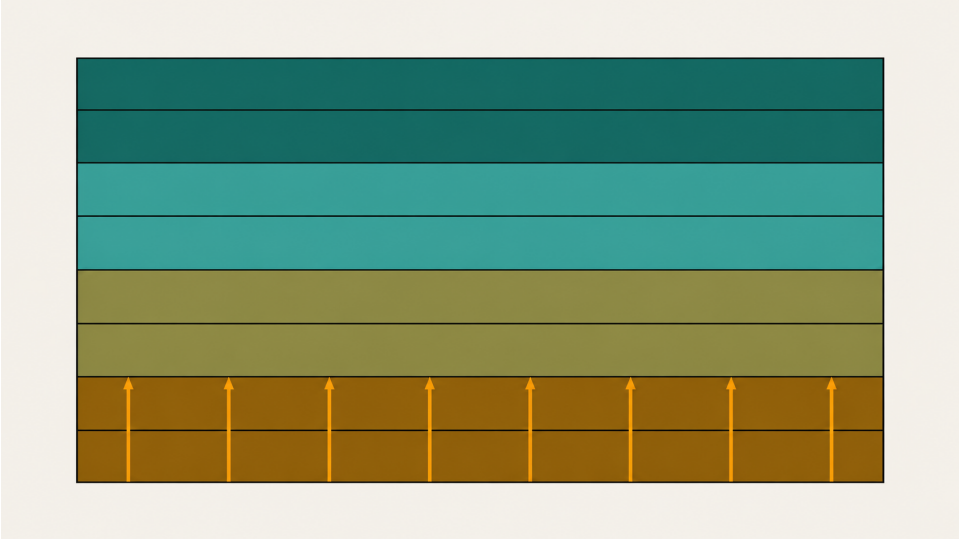
Sun and Wang arrive independently at our central structural criticism, expressed in the vocabulary of physics rather than of argument.

On τ scaling: it answers what to optimize but not where the limit lies, and it lacks an invariant, a quantity that stays fixed while everything else scales. They level the same charge at Moore's Law, which they call an empirical description anchored to nothing, comparing it to the astronomical tables that cataloged planetary motion for centuries before anyone could explain why the planets moved that way.

That is our "no arithmetic, no floor" finding, reached from the opposite direction by people qualified to reach it.

They then do something we could not: they put a number on the ceiling. Heat generated in the bottom layer of a stack has to travel up through the layers above it to escape, so the taller the stack, the harder it is to cool. That means a tall stack can be given less power than a short one, and the power it does get has to be shared among more transistors, so the speed at which it can run falls faster still.

FIG. 05 / THERMAL PATH THROUGH A TALLER STACK



Heat from buried layers must pass through upper layers to escape.

CONCEPTUAL SUPPLIED ARTWORK / NOT A MEASURED THERMAL PROFILE

Their conclusion, in their phrase, is that stacking eventually buys transistors that cannot all be switched on. Weighed against the wiring benefits that make stacking worthwhile in the first place, they place the optimum at roughly four to eight layers, with efficiency dropping sharply beyond it.

Set against Huawei's stated roadmap of three, four and more layers, that is a specific, quantified disagreement between two Chinese groups, on the record, ten days apart.

Where each side appears to be right, so far as we can judge it.

Sun and Wang's pessimism rests on one assumption. Their formula includes a term for how much energy it costs to move a single bit of data, and they treat that number as fixed by the manufacturing technology, something a chip designer inherits and cannot change. Every pessimistic conclusion follows from holding it constant while the stack grows taller.

But lowering exactly that number is the entire purpose of folding. Shorter wires cost less energy per bit; that is Huawei's whole mechanism. And Sun and Wang's own equation would register the improvement if the term were allowed to fall. They did not build a formula that forbids Huawei's result. They built one that assumes the designer has no lever on the variable Huawei is pulling.

Cutting the other way: Huawei is at two layers today, and the disputed territory begins around four. No measurement yet contradicts either party. It is also worth noting that Sun and Wang

were writing ten days before the block-level power data was published, so their model was built without the evidence that most directly challenges it.

We had reached an adjacent conclusion from an entirely different direction.

Our reasoning was about incentives and available space: that recovered area in a data center gets spent on computing rather than cooling, that a processor already at maximum manufacturable size has nowhere to spread heat sideways, and that the discipline which produced the cool phone result is unlikely to survive contact with a revenue-maximizing data center.

Their argument is about thermal resistance and arrives at a number. Theirs is the rigorous version, the kind only specialists can assert, and more useful than ours precisely because it names a height at which the question gets settled.

The strategic read

Invesco states the export-control consequence plainly and in an institutional register: with the most advanced lithography equipment blocked, China has moved to competing on architecture rather than transistor size, and τ scaling swaps transistor size for signal travel time as the measure of progress.

Their verdict is carefully balanced and, in our view, correctly placed. Unproven at scale, with heat still a hurdle, and the framing still matters, because a chip built on an older process could in principle rival a leading-edge one through architecture instead of miniaturization.

Their comparison table carries the whole strategic point in a single row.

Moore's Law depends on advanced lithography machines available only from the West; the τ approach is achievable on older equipment. And the bottleneck they name for τ, writing in July, is thermal management. Which means Huawei's September paper answers precisely the objection the investment community had already identified as decisive.

That sequence is telling: the thermal paper is not an academic aside, it is a response to the specific question gating the commercial thesis.

Two things in the Invesco note materially refine our findings rather than confirming them.

First, the geography of the power constraint.

We said power is the ceiling and that renaming progress moves no megawatts. Still true. But Invesco's data shows the ceiling sits at very different heights in different places: Chinese industrial electricity runs at a fraction of developed-market rates, national policy deliberately routes data centers toward regions with surplus power, and China dominates grid battery storage, which matters specifically because inference demand is intermittent. The constraint that bites hardest in Virginia leaves considerably less of a mark elsewhere. Cheap, buildable

power plus an efficiency-first hardware program is a coherent strategy even though neither repeals any law of physics.

Second, a partial answer to the gap we flagged as most important.

We argued that the missing design software is the real gate, and that nobody well placed to build it has a reason to. Invesco describes a very large state-directed data-center build-out carrying high domestic-sourcing requirements, alongside rules requiring data centers to source a substantial share of chips from domestic suppliers. A mandated captive market gives one ecosystem's design-software and packaging industries exactly the incentive we said was absent. It gives the other ecosystem nothing.

That converts our finding into a sharper and more consequential one. The tools will likely be built inside one ecosystem and not the other, which points toward the industry diverging into two toolchains optimizing different quantities, rather than converging on a new common measure.

One further point of coherence. Invesco's model-layer figures show Chinese AI models priced far below Western equivalents at comparable capability, with the whole domestic stack optimized for cost per unit of inference rather than for frontier capability. That is exactly what this hardware delivers: cheaper and quicker rather than smarter. The silicon strategy and the software strategy are the same strategy, which goes some way to explaining why this direction is being pursued as hard as it is by this particular actor.

Where the outside sources get things wrong

We applied the same instrument to both that we applied to ourselves above.

Invesco appears to describe the mechanism incorrectly, writing that unlike conventional stacking, which bonds separate chips together, Huawei folds circuitry within a single chip. Huawei's own description is of two wafers aligned, pressed together and heated until their surfaces fuse. It is bonded silicon, not a single monolithic piece. What differs is how fine the connections are and how small a unit of circuitry can be assigned across the boundary, not whether there is a boundary at all.

We note this without much criticism, since they were writing before the detailed process description was published. But the drift is worth flagging because of its direction: as the story travels it moves toward monolithic magic, and away from the thing that is actually novel and actually checkable, which is the fineness of the bonding.

Sun and Wang cite Huawei's claims more generously than Huawei's own measurements support, quoting density improvements of two to five times and performance gains of up to double, where the measured claim in the primary material is a single-generation density gain of roughly half. They are summarizing a launch announcement rather than the measurements.

We raise it gently, and mainly because of what it demonstrates: figures inflate in transit in both directions, in the skeptical account as well as the favorable one. That is an argument for going back to primary documents, which is most of what our method consists of.

What the overlap does and does not establish

Three independent readings, one from structural analysis, one from device physics, one from investment strategy, converge on four conclusions:

1. That the old measure of progress is exhausted
2. That τ scaling names a real target without yet being a usable measure
3. That the binding constraint is energy rather than latency; and
4. That value is migrating toward packaging, memory and integration.

Sun and Wang put the last of these most strongly, predicting that investment in packaging research will eventually exceed investment in front-end lithography.

What the overlap does not establish is foresight.

Both outside documents were published before we started researching — the first week of September 2026. Our findings were derived from the primary material without reference to them, which makes the convergence an independent check on method rather than a forecast.

We draw that line ourselves because a claim that cannot survive its own scrutiny is not worth making.

It is also worth saying that the two outside documents do not agree with each other. One argues the ceiling is physical and close. The other argues the framing matters commercially whether or not the ceiling is close. Both can be correct, and nothing in our reading requires choosing.

What to Watch

Layer count is the decisive test, and it has a schedule.

The whole disagreement reduces to a single number: how many sheets of working silicon can usefully be stacked. Sun and Wang derive a ceiling at four to eight, beyond which they say efficiency falls away sharply. Huawei ships two today and intends to go to three, four and more. Nothing is settled while Huawei stays below four, because their measurements and Sun and Wang's derivation are compatible in that range. The moment they pass it, one side starts being publicly wrong.

Bonding fineness is the real novelty and the most checkable claim.

Huawei reports a current figure of 1.5 micrometers between vertical connections, roughly 50 million of them across the chip. They say Kirin 2027 silicon already exists at 1 micrometer and over 100 million connections, and they target 720 nanometers and over 200 million within three years. For context, they characterize current advanced commercial practice as roughly 10 micrometers and under a million connections per 100 square millimeters. Each waypoint is dated, specific, and verifiable in a way that no projection about 2031 is. If the trajectory stalls, the claim stalls with it.

Watch whether three-dimensional design software appears, and from which ecosystem.

If it emerges from only one, the likely outcome is not adoption of a new industry standard but divergence into two incompatible ones.

Watch which way the dial gets turned in the data center.

The same folded silicon can be run in either of two settings: matched to its predecessor's performance while drawing much less power, or run flat out to deliver far more performance while drawing more power per square millimeter than the chip it replaced. Huawei chose the first for a phone, where battery life demands it. When they publish figures for a data-center part, look at which setting the figures describe. If the headline is throughput rather than efficiency, the thermal question that the September paper claims to have settled comes straight back, in the environment where it matters most.

Watch for the first independent measurement.

Everything to date is vendor-reported. A third-party teardown or benchmark of a shipping folded part would change the evidentiary picture more than any further paper from any party.

And watch the scoreboard.

No industry re-denominates until its benchmarks change. Today's still report a single number per workload. Until that changes, τ scaling remains an argument rather than a standard, however good the engineering underneath it turns out to be.

Methodology: SupraGraphos' Structural Realism Applied to Knowledge Mapping

We turn tacit human frameworks into explicit AI workflows. At the core of this approach is self-application: we abstract our own ways of working, mental models, and constantly changing (and hopefully improving) knowledge in Knowledge Representation and Reasoning, and transform our tacit lenses and cognitive processes into workflows and SOPs. Generative AI simply hastens the speed, improves the scope, and increases the scale of what we could otherwise do with pen on paper, whiteboards, or keyboards.

For this report, we aimed a set of 11 "Intuitions" towards a core list of sources. The Intuitions is a framework aligned with epistemic structural realism that treats natural language as a knowledge representation system that can be reasoned over directly. In plain language: it attempts to compose the actual structure underneath the words that are being used to convey them.

Applied here, the Intuitions work in combination rather than in sequence. Working from the Huawei material alone, we derived what the work is for, its central proposition, instead of accepting the authors' framing, and tested which claims would survive that proposition's removal. That is how the geometry argument separated from the τ theory.

We traced how the field's current measure of progress got installed and whether it was ever necessary, which led us to read the proposed new measure as a currency swap by an actor blocked in the old one.

We tested whether that measure holds together as a usable unit, and whether each demonstration supports the claim it is attached to.

We also set the demonstration, a handheld part drawing a few watts, against the claim, gigawatt facilities, since the two sit at very different scales of size, power and time.

Only once the September paper arrived did we weigh whether the incentive behind its favorable result travels with the technique, and that reading replaced an earlier, mistaken one on heat. Where two lenses disagreed, we treated the disagreement as a finding: one lens saw τ as a single unit repeating from transistor to data center, while another found that each layer runs on different mechanisms, so time saved in one layer cannot be traded for time in another.

Part Two tests overall findings, composed into a single read. Two outside documents, neither consulted while it was formed, were set against our conclusions to see where they converged, diverged or refined them. In two places they refined them, and we say where.

Sources

Primary

- He, T. "A Time Scaling Theory for Multi-Layer Electronic Systems." ChinaXiv 202605.00224v2, posted 3 July 2026. Cited in the September paper as Science China Information Sciences, vol. 69 no. 8, August 2026.
- He, T. "Huawei's τ Chip Was Supposed to Melt?" ChinaXiv 202609.00031v1, posted 4 September 2026.
- Huawei. "HUAWEI Presents the Tau (τ) Scaling Law." Press release, 25 May 2026.

Reports

- Sun, B. H. and Wang, E. X. "From Moore's Law to the Volume-Inverse Law." Preprints.org, doi:10.20944/preprints202608.1596.v1, 24 August 2026. Not peer reviewed.
- Oerth, A. and Wu, T. "An alternative model: China's low-cost AI strategy." Invesco Strategy & Insights, July 2026. Marketing communication for professional investors.

Press

- South China Morning Post, "Huawei paper shows Tau Scaling Law chip solves overheating ahead of Kirin 2026 launch," 7 September 2026.
- Interesting Engineering, "Huawei's new chip design tackles overheating with 55% higher transistor density," 7 September 2026.

Context figures not drawn from the above (grid queues and data-center demand; accelerator bill-of-materials composition; memory-bandwidth limits on token generation) are trade-press and industry estimates rather than audited disclosures, and are used here as approximate context only.



About SupraGraphos

SupraGraphos is a research and advisory consultancy focused on structural market analysis and knowledge infrastructure. The analysis in this brief was produced using our proprietary analytical methodology, built to surface the underlying drivers behind market pricing and behavior — not just summarize the surface-level narrative.

ENGAGEMENTS

SupraGraphos works with organizations that need to make high-stakes decisions in complex, fast-changing markets. Engagements include market and structural analysis, knowledge architecture, GenAI implementation, and custom automation pipelines for mission-critical workflows. Early clients are engaged under marketvalidation terms.

Contact: **gino[@]supragraphos.com**

Prepared by Gino R. Diño for SupraGraphos