A time scaling theory for multi-layer electronic systems

Tingbo He[*]

Huawei Technologies Co., Ltd.

Abstract

For six decades, Moore's geometric scaling drove progress in semiconductors. That industry compact no longer holds: returns from pure dimensional shrinking have flattened, leading-edge design budgets exceed one billion dollars per chip, and cost-per-transistor at the most advanced nodes is no longer falling. This perspective argues for a successor scaling principle — τ scaling — that adopts time itself, rather than transistor area, as the primary metric of progress, applying a single characteristic time constant τ as the unifying optimization target across twelve orders of magnitude, from a switching transistor to a data-center workload. Two production-scale demonstrations are presented. On a mobile SoC, *LogicFolding* — a methodology that partitions digital, analog, and memory circuits across vertically stacked active tiers — delivers a 55% step-wise increase in transistor density and a 41% reduction in power consumption at equivalent performance at a fixed device node. On AI systems, a co-designed stack comprising the memory-semantic Unified Bus fabric, near-packaged Hi-ONE optical I/O, and edge-to-surface 3D Folding projects more than 100× growth in hardware integration by 2035.The deeper claim is methodological: τ scaling is the first scaling principle since Dennard to establish a shared optimization target across the entire computing stack.

Keywords τ scaling, LogicFolding, gear ratio, wafer-to-wafer hybrid bonding, Unified Bus, Hi-ONE

* Corresponding author (email: hetb@huawei.com)

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

ChinaXiv:202605.00224v2

1 Introduction

Since the mid-1960s, the semiconductor industry has measured progress in nanometers. Every eighteen months, transistors shrank, frequencies rose, and the cost per logic gate fell. Moore's Law functioned as both an empirical observation and helped establish an industry compact upon which the entire computing stack was built. That industry compact no longer holds. Beyond the 7 nm node, geometric scaling no longer delivers its historical dividends. Lithography tooling is approaching the physical limits of patterning, EUV depreciation dominates wafer cost, and the per-transistor price curve has flattened — and in some cases reversed. For organizations whose access to the most advanced lithography is constrained, the constraint became binding earlier and bears down more severely.

The central question for the industry has therefore changed. It is no longer *"how much further can the transistor shrink?"* It is *"what should be scaled, and against what objective?"*

Over the past six years, the author's team at Huawei Semiconductor has investigated this question in silicon across mobile SoCs, AI accelerators, system fabrics, and packaging. The conclusion is that the answer lies not in another node, nor in another transistor architecture, but in a change of the primary optimization target itself. This perspective argues that the next decade of electronic-system evolution should be guided not by geometric scaling, but by time scaling — the systematic reduction of a single characteristic time constant τ across every layer of the stack, from a transistor switching in a picosecond to a data-center workload responding in a second.

The case for τ scaling is developed below as both a scientific methodology and an industrial roadmap, drawing on lessons from 381 chips brought to volume production between May 2020 and May 2026.

2 The end of the geometric era

For most of its history, the semiconductor industry has had one job: make the transistor smaller. Gordon Moore's 1965 observation — that transistor density doubles approximately every two

This version posted 2026-07-03.

years— was complemented a decade later by Robert Dennard's scaling theory, which established that proportional shrinking of voltage and dimensions could maintain a constant electric field [1, 2]. Together, geometric scaling and Dennard scaling delivered exponential improvements in performance per watt and performance per dollar for nearly five decades.

This arrangement unraveled in two stages. Around 2005, Dennard scaling broke first: voltage ceased to scale proportionally with feature size, and the dark-silicon era began. Geometric scaling persisted longer, sustained by FinFET and subsequently gate-all-around (GAA) device architectures. Beyond 7nm, however, returns from pure dimensional scaling have flattened. The reasons are now well documented: velocity saturation reduces the dependence of intrinsic delay on channel length from quadratic to linear; the parasitic resistance and capacitance of local interconnects increasingly dominate the standard-cell delay budget; mask costs, EUV depreciation, and design-rule complexity have driven leading-edge chip design budgets past one billion dollars per chip at the 2 nm node [3-8].

The economic consequences are equally inescapable. Cost per transistor has flattened at advanced nodes and, at the leading edge, is now rising. The industry compact that sustained the last fifty years — *more transistors at lower cost every generation* — no longer holds.

For Huawei Semiconductor, this transition arrived with an additional constraint: restricted access to the most advanced lithography tooling. Assuming that another node would resolve the problem was no longer tenable. Six years ago, the geometric roadmap plateaued, forcing a more fundamental question — one that, in retrospect, the entire industry will eventually have to confront.

3 Time, not space: the real currency of Moore's era

Reduced to its essential effect on the end user, Moore's Law was never fundamentally about geometry. Smaller transistors improved system performance because they switched faster. Denser interconnects improved performance because signals traversed shorter distances. Higher integration improved performance because data crossed fewer boundaries. What each generation

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

delivered, in essence, was a reduction in time — picosecond to nanosecond at the device, nanosecond to microsecond at the chip, microsecond to second at the system. Spatial scaling served merely as the instrument for compressing time.

Once this is recognized, an obvious reframing presents itself. Time itself should be adopted as the primary metric. A characteristic time constant τ can be defined at every layer of the stack — transistor, circuit, chip, and system — and its reduction treated as the unifying optimization target. Geometric scaling then becomes one technique among many for reducing τ, rather than the only one.

This principle is called τ scaling, and is proposed here as the successor to geometric Moore scaling as the guiding principle of semiconductor evolution. Formally, τ is treated as a layered construct that decomposes as

$$\tau = f(\tau_{transistor}, \tau_{circuit}, \tau_{chip}, \tau_{system}),$$

where $\tau_{transistor}$, $\tau_{circuit}$, τ_{chip}, and τ_{system} represent the time constants at the transistor, circuit, chip, and system layer, respectively. Each layer's τ composed from the layers beneath it together with the organizational and communication overheads introduced at that layer. As illustrated in Figure 1, the working space of τ spans approximately twelve orders of magnitude in time (picoseconds to seconds) and a comparable range in space (nanometers to kilometers). At each layer, distinct mechanisms are available for reducing τ:

- Transistor: intrinsic switching delay, addressed through mobility enhancement, strain engineering, high-κ/metal gate, and GAA architectures, and, increasingly, through reduction of the parasitic R and C of local interconnects, which now exceed the intrinsic transit time by several factors [6, 7].
- Circuit: RC propagation delay along signal paths, addressed through lower-resistivity conductors, low-κ dielectrics, and — most consequentially — through reduction of wire length via vertical integration [9, 10].
- Chip: compute and memory-access latency, addressed through architectural choices, pipeline

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

depth, memory hierarchy, and on-chip fabrics [11].

- System: end-to-end message and synchronization time, addressed through interconnect topology, protocol stack, and fabric design [12].

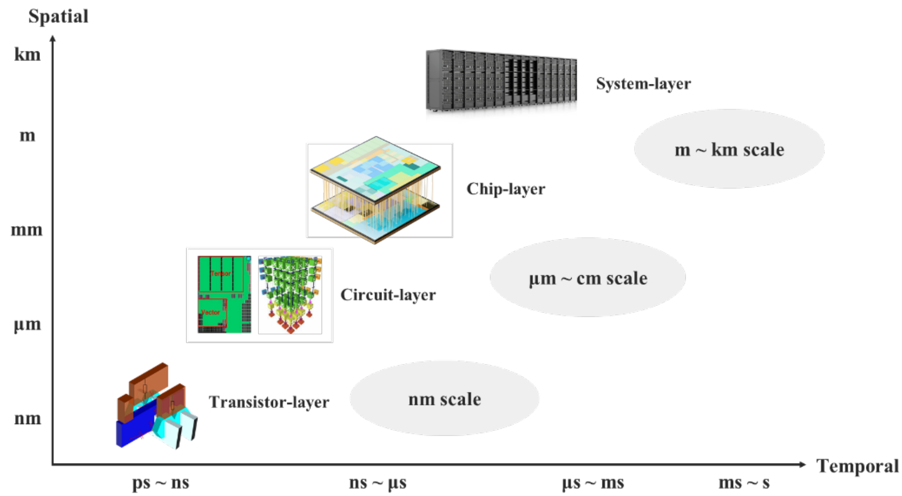

Figure 1. The working space of τ scaling spans 12 orders of magnitude across both the temporal and spatial dimensions and is partitioned into four layers: transistors, circuits, chips and systems.

A useful generational rule emerges from this layered formulation:

$$\tau_{n+1} = \frac{\tau_n}{\alpha},$$

where the subscripts n and n+1 respectively denote the current and subsequent generations. According to the differing market pressures and optimization priorities across diverse industry sectors, we present that the annual scaling factor is application-specific rather than universal. Different industry sectors demand distinct acceleration factors driven by unique application constraints. We project that the annual scaling factor over the next decade will be approximately 1.3 for mobile devices constrained by power and thermal budgets, 1.5 for autonomous driving systems requiring safety-critical real-time response, and up to 10 for artificial intelligence (AI) token generation where throughput directly translates to economic value.

What renders τ a useful primary metric, rather than a relabeling of existing ones, is that it is the

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

same metric across the entire stack. Frequency, latency, bandwidth, and throughput are all governed by τ at their respective layers. A process technologist, a circuit designer, and a system architect can debate the same quantity in identical units. τ is the language that enables end-to-end stack co-optimization — and the era of independent optimization at each layer, with timing emerging as a residual, has concluded.

4 *LogicFolding*: a mobile-SoC proof point

The first production-scale test of τ scaling was conducted in mobile. A smartphone SoC is the unusual case in which one chip constitutes the entire system. Multi-socket parallelism is not available; no thousand-node fabric can mask a slow link. All performance delivered to the user originates from a single die, under a few-watt power envelope, against thermal limits set by handheld form-factor constraints.

After 2020, when access to leading-edge nodes was restricted, the operative question became: with the node fixed, how can generation-over-generation improvements continue to be delivered on a single die?

The answer that emerged is called *LogicFolding*.

Definition. *LogicFolding* is a design methodology that partitions digital, analog, and memory circuits across vertically stacked active tiers to jointly optimize performance, power, and area following the time scaling principle (Figure 2).

Digital circuits divide into combinational logic — the Boolean network between registers — and sequential logic — the flip-flops that hold state. The performance ceiling of a digital system is set by the critical-path delay between adjacent flip-flop stages, which in turn is dominated by interconnect RC and gate count along that path. Conventional optimization places gates in a plane and routes wires through a metal stack above; the longer the wire, the greater the parasitic RC, and the slower the critical path.

LogicFolding abandons the planar assumption. Critical-path gates are distributed across two (and

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

eventually more) vertically stacked active tiers, connected through ultra-fine-pitch hybrid bonding. From the circuit designer's perspective, the two tiers behave as a single continuous fabric, with cells distributed across the wafer boundary as if it were an additional metal layer. Signal wires become substantially shorter, parasitic RC decreases sharply, clock skew tightens, and the chip operates at a higher clock frequency at the same device node.

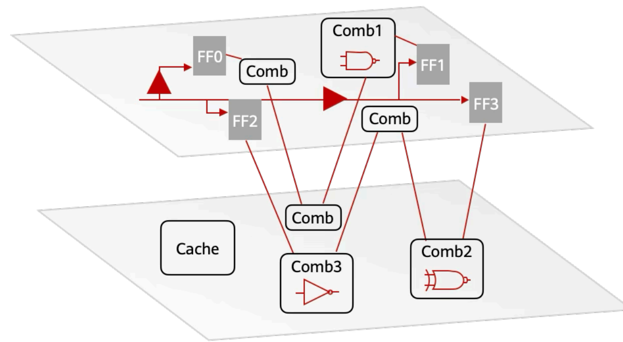

Figure 2. The schematic illustration of *LogicFolding*.

To fully realize the architectural benefits of *LogicFolding*, it is critical to maintain a low pitch ratio (often referred to as the "gear ratio") between the hybrid-bonding and the top metal routing layer. As the vertical interconnect pitch approaches the dimensions of the top metal layer, the nature of the optimization objective undergoes a fundamental transformation. Historically, when the vertical interconnect pitch is much sparser than the top metal pitch, the design space is fundamentally restricted to a discrete optimization problem. Designers manually defined partition boundaries at the macro level, assigning entire functional blocks to specific dies [13-18]. The coarse granularity of inter-die connections forced a discrete block assignment methodology, which is computationally tractable but not globally optimal. *LogicFolding* proposed here is positioned as a continuous optimization problem, in which fine-grained vertical integration enables the design space to be explored at a level of resolution much higher than that of functional blocks, opening the door to globally coordinated optimization of circuits across the vertical dimension. As vertical interconnect density increases through the progressive shrinking of bond pad pitches, wafers are effectively brought together ever closer from a circuit connectivity perspective. This enables the optimization

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

space transitions from discrete to continuous, requiring the use of advanced automated design tools. It is worth noting that while sequential 3D integration theoretically offers the ultimate fine-grained device or standard-cell granularity by fabricating device layers sequentially, it currently faces significant manufacturing bottlenecks [19-22]. Most critically, the lower-layer devices are highly prone to performance degradation due to the strict thermal budget constraints inherent in the sequential fabrication process. As a commercially viable realization, *LogicFolding* achieves the necessary low gear ratio for continuous optimization by utilizing mature, advanced wafer-to-wafer hybrid bonding.

In practice, *LogicFolding* requires the gear ratio to fall below approximately 3, with lower ratios generally better. With today's top-metal pitch around 720 nm, this translates into a hybrid-bonding pitch below 2 μm — and ideally to a gear ratio of approximately 1, at which the bird-cage routing overhead at the bonding interface effectively vanishes. Achieving this pitch, together with the required overlay accuracy (<0.5 μm), TSV scaling (CD and KOZ sub-1.5 μm, pitchsub-6 μm), and yield (~100% with smart redundancy), required a multi-year process-development effort across the supplier and partner ecosystem.

The results measured on Kirin 2026 yield concrete practical evidence when compared to the 2025 Kirin9030 Pro baseline. Although both are fabricated on the identical mature process node, the baseline employs a traditional planar design whereas Kirin 2026 utilizes *LogicFolding*:

- Transistor density rose step-wise from 155 to 238 MTr/mm² in a single generation (transistor density is calculated using the formula $\frac{2}{CPP \times cell\ height}$; the area utilization of Kirin SoC design is 68%) — a magnitude of improvement that previously required three years of geometric scaling.
- The maximum clock frequency of SoC performance-core rose by nearly 13% at an ambient temperature with a 1.1V supply voltage.
- A high-speed global Network-on-Chip data path constructed across both upper and lower tiers reduced the data-path footprint by 55%, with improved power-delivery stability.

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

- A post-silicon clock-skew adjustment scheme contributed over 5% SoC performance independently.
- On SRAM — where access speed, energy-per-bit, and area depend strongly on bit-line and word-line length — *LogicFolding* shortened critical paths, reduced energy per bit, and increased operating frequency by over 40%.
- On a representative processing core, the double-layer folding architecture reduced clock-buffer count by more than 50%, clock skew by 25%, and wire length by approximately 30%.

Thermal management remains the critical challenge in the *LogicFolding* architecture. To address this, we employ a thermal-aware partitioning and floorplanning strategies. During the design phase, we deliberately avoid folding high-power circuits and structurally prevent the spatial adjacency of high-power subsystems. The SoC performance-core serves as the focal point of our detailed evaluation. As shown in Table 1, leveraging the performance gains enabled by *LogicFolding*, Kirin 2026 lowers its supply voltage to achieve iso-performance with the Kirin9030 Pro. Consequently, practical measurements at this iso-performance target reveal a 41% reduction in power consumption alongside a 5.6% decrease in power density.

Table 1. The comparison of power at iso-performance between Kirin 2026 and Kirin9030 pro

	Kirin9030 Pro	**Kirin2026**
Temperature	25 °C	25 °C
Voltage	1.1 V	0.9 V
Frequency	2.75 GHz	2.5 GHz
Normalized power at iso-performance	1	0.59
Normalized area	1	0.625
Normalized power density	1	0.944

These gains were achieved at a *fixed* device node, obtained not through a new lithography step but through a topological reorganization of the spatial distribution of logic in three dimensions.

The *LogicFolding* implementation shipping in Kirin 2026 is deliberately conservative. The hybrid-bonding pitch reached 1.5 μm; TSV landing advanced only one step below the top metal; folding was applied selectively along key critical paths rather than across the entire design (Figure 3). Even so, the CPU performance-core frequency returns to 3.1 GHz this year.

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

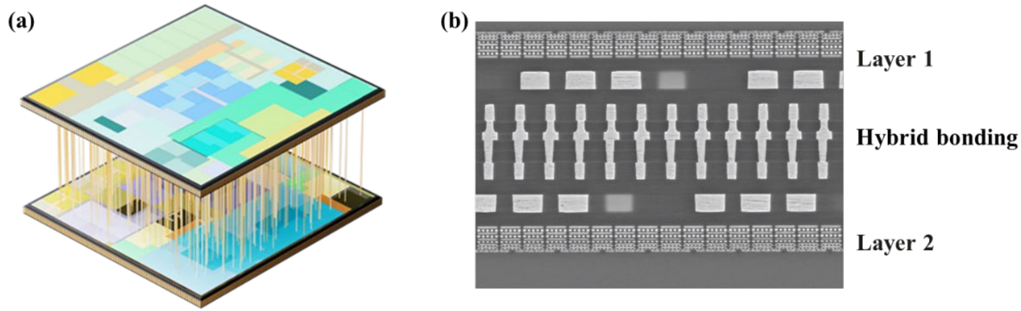

Figure 3. (a) The schematic illustration of the next-generation Kirin SoC platform; (b) Cross-section image of its bonding interface.

Over the next decade, *LogicFolding* is expected to evolve from local critical-path folding to full-scale, multi-layer folding — three, four, and more active tiers per package — enabled by lower-temperature hybrid bonding (relaxing the thermal budget across tiers) and by TSV landing migrating from the top metal down to M6, which liberates over 30% of high-level routing resources. From 2026 to 2035, transistor density is projected to rise toward 400 MTr/mm^2 and beyond. Simultaneously, *LogicFolding* enables Kirin to substantially step up CPU core frequency, and paves the ways towards 4 GHz and beyond (see Figure 4 and Table 2). The roadmap is feasible and, in cost terms, economically viable.

Table 2. Trend of the operating frequency of Kirin CPU performance core.

	SoC	**Architecture**	**Frequency**	**State**
2023	Kirin9000s	Planar	2.6 GHz	Mass product
2024	Kirin9020	Planar	2.65 GHz	Mass product
2025	Kirin9030 pro	Planar	2.75 GHz	Mass product
2026	Kirin 2026	LogicFolding	3.1 GHz	Silicon
2027	Kirin 2027	LogicFolding	3.39 GHz	Silicon
2028	Kirin 2028	LogicFolding	3.71 GHz	Pre-silicon
2029	Kirin 2029	LogicFolding	4 GHz	Pre-silicon

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

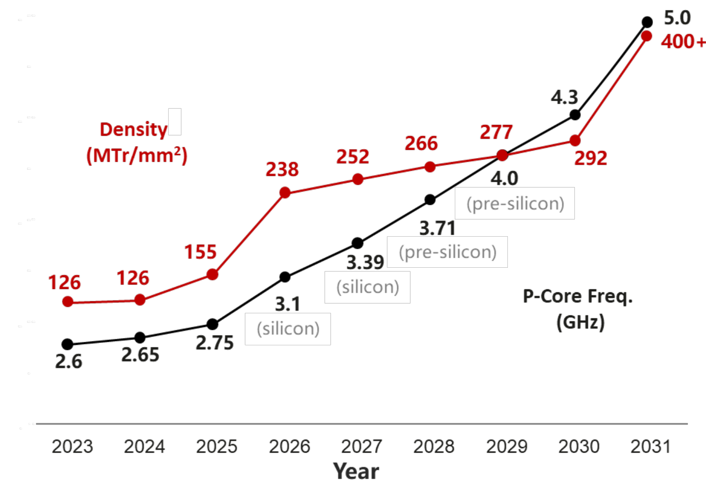

Figure 4. Projected transistor density and performance-core frequency for our future Kirin products.

Highlight — *LogicFolding* at a glance

- Hybrid-bonding pitch: sub-2 μm (1.5 μm in Kirin 2026; target gear ratio ≈ 1)
- Overlay accuracy: under 0.5 μm
- TSV CD/KOZ: sub-1.5 μm; pitch sub-6 μm; failure rate <100 ppm; repair rate 99.9%
- Yield: ~100% with smart redundancy
- Transistor density: 155 →238 MTr/mm² in a single step
- Power-efficiency / frequency gain (SoC P-core): +41% / +13%
- SRAM operating frequency: +40%+
- Clock-buffer count / clock skew / wire length on a representative core: −50% / −25% /−30%

5 From picoseconds to microseconds: τ scaling in the AI data center

A natural question is whether a principle developed in the milliwatt smartphone regime survives translation to the gigawatt regime of AI training and inference. AI workloads occupy the opposite end of the τ spectrum: not a single chip but hundreds or thousands of chips behaving as one machine, with aggregate compute increasing by approximately six orders of magnitude over the past decade. The answer is affirmative — provided τ is treated as a system-level objective and applied across

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

the whole chain, rather than within a single accelerator.

Two facts shape the AI side of the τ argument. First, AI systems continue to grow — from one chip, to dozens, to hundreds, and increasingly to tens of thousands [23, 24]. Second, the energy budget and the materials budget of modern AI systems are dominated by data, not by compute [25-27]. Over 80% of energy in a large AI cluster is consumed by data movement; over 70% of system cost is allocated to data storage. The implication is direct: reducing the time data spends *in transit* — between chips, between racks, and within the package — is at least as important as reducing the time compute spends *computing*.

τ scaling is instantiated at AI scale through three coordinated layers: a system fabric (Unified Bus), a near-packaged optical engine (Hi-ONE), and a topological reorganization of the package itself (3D Folding). This full-stack approach systematically compresses the system τ inherent to distributed AI systems. Specifically, the Unified Bus (UB) eliminates multi-layer protocol overhead via unified memory-semantic interconnects, drastically reducing cross-node communication latency. Hi-ONE leverages near-packaged optical I/O to directly compress physical transmission delay. 3D Folding overcomes the quadratic-to-linear scaling divergence by relocating the edge-bound resources onto surfaces, minimizing intra-node parasitic RC. Collectively, this circuit-to-system optimization stack realizes the ultimate target of τ scaling in AI system: empowering the large-scale AI cluster to operate cohesively as a single logical entity.

5.1 Unified Bus — a τ-first system fabric

Traditional multi-node, multi-accelerator architectures move data across multiple stacked protocols: PCIe to the host, NVLink or proprietary fabrics within the chassis, Ethernet or InfiniBand between chassis, and software-stack remote-memory access on top. Each layer entails a protocol conversion, additional serialization, an extra DMA buffer, and a further handshake. Every conversion adds latency, reduces reliability, and incurs additional cost.

Unified Bus (UB) replaces this stack with a single protocol that operates within *and* across the chassis — a fully peer-to-peer fabric that exposes memory semantics natively across the whole

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

system (Figure 5). Data movement is reduced to conversion-free, peer-to-peer transmission at the memory-semantic layer, with hardware-managed coherence in place of software-stack message passing.

The measured benefit is approximately two orders of magnitude: end-to-end remote-access latency falls from the tens of microseconds typical of TCP/IP-class stacks to approximately 100 ns — a ~500× reduction in system τ along the dominant communication axis [28-30]. At the rack scale, this brings the system asymptotically close to a single, fabric-coherent machine — designated internally as a *System-as-One-Chip*.

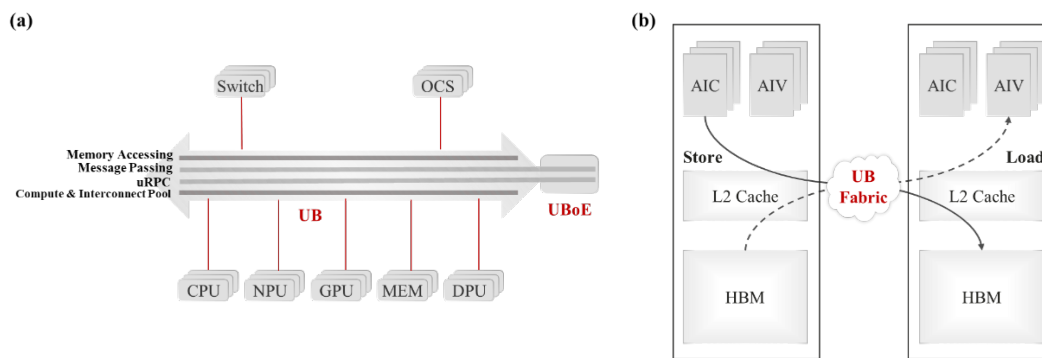

Figure 5. (a) The Unified Bus (UB) natively supports memory access semantics, message passing, and unified remote procedure call (uRPC). This enables the seamless integration of diverse computational resources, achieving high bandwidth and ultra-low latency while facilitating efficient resource pooling across the distributed system; (b) Schematic illustration of low overhead memory access via UB.

5.2 Hi-ONE — optical I/O at the package

Once communication latency is reduced, the next bottleneck shifts. Increasing the density of chips within a single rack pushes power density and reliability past their limits — and pushes electrical SerDes past theirs. At 400 Gb/s per AI chip, copper cabling remains well understood and reliable. At multi-Tb/s per chip, copper becomes physically impractical: SerDes reach contracts, cabling becomes prohibitively bulky, panel installation becomes infeasible, and thermal and power-

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

delivery margins are exhausted [31, 32].

The approach developed at Huawei Semiconductor is the High-density Optical-interconnect-Node Engine, Hi-ONE — a near-packaged optical engine that delivers 8 Tb/s per module, matching the UB bandwidth of an AI chip on a single optical link (Figure 6). It reduces the required SerDes reach from ~100cm to ~5 cm, eliminates bulky cabling, and extends reach from under a meter to 100 meters —rendering high-density interconnect for distributed, gigawatt-scale data centers physically realizable.

The design philosophy underlying Hi-ONE is itself a τ-scaling argument. In place of a heavy DSP for high signal fidelity, Hi-ONE adopts a linear approach — an analog equalization-enhanced driver and trans-impedance amplifier — and permits the UB protocol to tolerate a deliberately relaxed bit-error rate. This cross-layer trade between protocol layer and physical layer reduces power, cost, and integration complexity, and epitomizes the cross-layer trade-off that a τ-first methodology rewards.

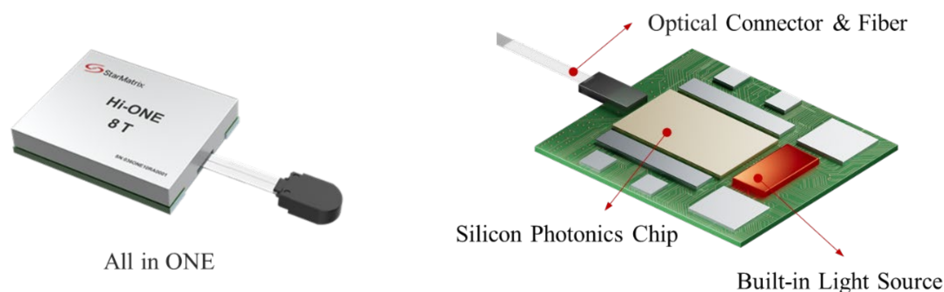

Figure 6. The illustrated image of a Hi-ONE chip.

5.3 The N²-vs-N dilemma, and why 3D Folding is inevitable

The deepest reason AI accelerators will not stop at 2.5D fan-out is geometric, and merits explicit statement because it determines the post-2030 roadmap.

In a conventional 2.5D AI chip, the logic die occupies the center of the package, HBM stacks and SerDes line its edges, and voltage regulators surround the package. Every memory signal, every interconnect signal, and every ampere of supply current must traverse the die's edge to reach the

ChinaXiv:202605.00224v2

compute resources within. If the die has side length N, then:

- compute capacity scales as N^2 (area),
- but memory bandwidth, interconnect, and power delivery — all carried by the 2.5D fan-out along the edge — scale only as N (perimeter).

The widening divergence between these quadratic and linear curves constitutes the fan-out dilemma, and it accounts for the stalling of 2.5D scaling independent of how aggressive the underlying logic node becomes. No transistor-level improvement closes a topological deficit.

3D Folding resolves this dilemma by relocating the edge-bound resources onto surfaces. Power delivery (via backside power and integrated voltage regulators), high-speed memory (via hybrid bonding to logic), and optical I/O (via near-packaged Hi-ONE) all migrate from perimeter to vertical surface — and, once located on a surface, they scale as N^2, matching the quadratic pace of compute. The package is no longer a logic die surrounded by a perimeter belt of memory and SerDes; it becomes a vertically integrated stack in which memory, fabric, power, and logic all scale together.

The roadmap places this evolution on an explicit timeline. Through approximately 2030, AI accelerators (the Ascend SuperPoD line — Ascend 910C in 2025, Ascend 950 in 2026, and the 990to follow) rely on a combination of mature techniques: chiplets, 2.5D fan-out, and 3D stacking via micro-bump and standard-pitch hybrid bonding. Around 2030, Ascend 990 will introduce *LogicFolding* into the AI accelerator class, and from that point 3D Folding becomes the principal carrier of α through 2035. Along this path, hardware integration is projected to increase by more than100× by 2035, with τ reduction distributed across every layer of the stack rather than concentrated at the device level.

Highlight — τ at AI system scale

- UB remote-access latency: ~10s of μs → ~100 ns (≈ 500× τ reduction)
- Hi-ONE per-module bandwidth: 8 Tb/s (matches per-chip UB bandwidth)

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

- Hi-ONE SerDes reach: ~100 cm → ~5 cm; panel-to-panel reach: <1 m → 100 m
- Fan-out dilemma: compute $\propto N^2$, perimeter-bound BW/I/O/power $\propto N$
- 3D Folding: relocates BW, optical I/O, and power delivery from edges onto surfaces, restoring

N^2 parity

- 2026 → 2035 projected hardware-integration growth: >100×

6 Logic and memory: from decoupling to re-fusion

One implication of τ scaling warrants separate discussion, because its consequences are industrial as well as technical.

In the 8086 era, the industry deliberately *decoupled* processors and memory through standardized memory buses. That decoupling permitted two industries to scale independently: processor performance advanced rapidly along the Moore curve, while memory vendors developed a vast, separate market alongside it.

The AI era is reversing this decoupling. The continuing expansion of compute density is pushing memory bandwidth, latency, power, and packaging to their limits. HBM, hybrid bonding, and 3D-stacked SRAM are symptoms of a single underlying fact: for modern AI workloads, data movement is as critical as computation itself, and logic and memory are once again being driven into tight physical integration. As they fuse, the balance of influence in the supply chain is shifting toward memory and packaging vendors.

The technological direction is unambiguous, but the economic resolution is not yet settled. Enduring success in the AI hardware era will accrue to those who can fuse logic and memory technologically *and* establish an economic partnership that allows both industries to share the benefits of that fusion over the long term. This is not merely a research problem; it is a structural problem for the industry to address over the next decade. By rendering the cross-layer cost of every separation visible, τ scaling ensures that the problem cannot be deferred.

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

7 Open challenges

It would be misleading to present τ scaling as a completed system. Several substantive problems remain open, and are identified here both to highlight ongoing work and to invite collaboration.

Toolchains and methodologies. Today's EDA was developed for an era in which area, timing, and power were optimized along three separate axes, with system τ emerging as a residual. Full-scale *LogicFolding* requires the toolchain to treat multiple stacked dies as a single continuous design entity — partitioning logic at cell granularity rather than block granularity, placing across the full volume under a unified cost function, and performing timing closure across inter-die paths where vertical-interconnect parasitics, KOZ exclusions, and inter-wafer process variation interact in ways that traditional 2D-trained tools do not address adequately. Preliminary internal tools have been developed that produce useful results, and methodology details will be published in the coming months. A τ-native toolchain — open, multi-physics, and 3D-native — is the single most important enabling investment for the next decade.

Inter-wafer process variation. *LogicFolding* bonds wafers from potentially distinct lots — and in some cases distinct nodes. Inter-wafer variation in V_{th}, drive current, and interconnect RC is materially greater than within-wafer variation, and falls most heavily on clock distribution and hold-time margins. Smart redundancy, adaptive compensation, and τ-aware signoff flows are necessary components of the response.

Vertical-interconnect overhead. Every hybrid bond and every TSV incurs a finite resistance and capacitance penalty, and TSV KOZ displaces standard cells. *LogicFolding* must therefore be justified layer by layer through the simple inequality

$$\tau_{Benefit}\ (effective\ silicon\ area\ +\ wire\ length\ reduction)\ >$$
$$\tau_{Penalty}\ (vertical\ interconnect\ RC).$$

This threshold has been crossed for mobile critical paths and for memory; the threshold is workload-specific, and the boundary will move as bonding pitch shrinks.

Energy. τ is a time law, not a joule law. A super-node operating 10× faster but with 10× greater

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

power consumption violates no scaling principle, yet exceeds grid capacity. τ scaling therefore requires an *energy companion*: memory-semantic fabrics that eliminate stack overhead, near-/co-packaged optics that reduce picojoules per bit by orders of magnitude, backside power delivery, compute-in/near-memory, and the disciplined practice of trading τ headroom back for power (DVFS at data-center scale —the same mechanism that enabled smartphone battery longevity). Importantly, τ headroom itself provides energy headroom when allocated in that direction.

Benchmarks. The industry's current performance benchmarks — Linpack, MLPerf, SPEC — were designed for an era in which a single scalar per workload sufficed. A τ-scaling industry requires τ-profile benchmarks — vectors that expose the dominant τ at each layer of a system together with the headroom remaining at that layer. The dominant-τ layer is, by definition, the next investment.

8 Six years in, ten years out

Between May 2020 and May 2026, Huawei Semiconductor designed and brought to volume production 381 chips serving mobile, AI, automotive, industrial, and infrastructure markets. Across that portfolio, the τ scaling thesis has held up:

- At the device and circuit layers, transistor density has risen from 155 toward 400+MTr/mm² by 2031.
- At the chip layer, *LogicFolding* has demonstrated, on a leading-edge mobile SoC, that critical-path frequency, power efficiency, and density can continue to advance at a fixed device node.
- At the system layer, Unified Bus and Hi-ONE have demonstrated that hundreds of microseconds of communication τ can be compressed to hundreds of nanoseconds, and that a multi-rack AI cluster can behave as a single coherent machine.
- Looking forward, CPU performance-core frequency is expected towards 4 GHz and beyond by 2029, Kirin SoC efficiency is projected to more than double in three to five years under typical use, and AI hardware integration is expected to grow more than 100× by 2035.

The deeper claim, beyond any individual product, is methodological. τ scaling is the first scaling

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

principle since Dennard to give the entire stack a shared optimization target. It signals to process technologists, circuit designers, architects, system engineers, and software teams that these communities are now optimizing the same quantity in identical units, and that improvements at any single layer must propagate to the system τ to count. It also indicates to industry strategists and capital allocators that the next dollar should follow τ, not nodes — that competitive performance no longer requires perpetual residence on the leading edge of lithography, and that packaging, memory bandwidth, and fabric design now command the strategic weight that the leading-edge logic node alone previously held.

For a generation of engineers educated to treat "Moore's Law" as synonymous with "progress," this is a difficult transition. The geometric era has, in fact, concluded; denial of that fact is not a viable strategy. The era of acceleration through miniaturization is giving way to an era of acceleration through τ optimization across the multi-layered electronic system — and the companies, research groups, and ecosystems that adopt τ as the primary objective in the next six to ten years will determine the shape of computing in the decade thereafter.

The next ten years of work are scoped. Many open questions remain, and no single organization can address them alone — the toolchain, the standards, the benchmarks, the device physics, and the economic models all require contributions from beyond any one company. This perspective is therefore intended as both a report from the field and an invitation.

The roadmap ahead is demanding, but the direction is unambiguous.

Author

Tingbo He leads Huawei's semiconductor business. The team she directs has designed and brought to volume production 381 chips between 2020 and 2026 across mobile, AI, automotive, and infrastructure markets, and is the source of the τ scaling methodology and the *LogicFolding*, UnifiedBus, and Hi-ONE technologies described in this article.

ChinaXiv:202605.00224v2

This version posted 2026-07-03.

ChinaXiv:202605.00224v2

Acknowledgments

This perspective draws on six years of work by thousands of engineers across Huawei Semiconductor and its ecosystem of foundry, equipment, EDA, and system partners. The author thanks the customers whose patience made this work possible.

References

1. Moore G E. Cramming more components onto integrated circuits. Electronics, 1965, 38: 114-117.

2. Dennard R H, Gaensslen F H, Yu H N, et al. Design of ion-implanted MOSFET's with very small physical dimensions. IEEE Journal of Solid-State Circuits, 1974, 9: 256-268.

3. Wong H S P. Beyond the conventional transistor. IBM J Res Dev, 2002, 46: 133-168.

4. Zhirnov V V, Cavin R K, Hutchby J A, et al. Limits to binary logic switch scaling - a gedanken model. Proceedings of the IEEE, 2003, 91: 1934-1939.

5. Moore G E. Lithography and the future of Moore's law. Proc SPIE, 1995, 2437: 2-17.

6. Taur Y, Ning T H. Fundamentals of Modern VLSI Devices. 3rd ed. Cambridge: Cambridge University Press, 2021.

7. IEEE. International roadmap for devices and systems (IRDS) 2023/2024: Interconnect and More-than-Moore chapters[R]. Piscataway: IEEE, 2024.

8. Flamm K. Measuring Moore's law: evidence from price, cost, and quality indexes[R]. Cambridge: National Bureau of Economic Research, 2018.

9. Soulié J P, Sankaran K, Troeye B V, et al. Selecting alternative metals for advanced interconnects. J Appl Phys, 2024, 136: 171101.

10. Batude P, Ernst T, Arcamone J, et al. 3D sequential integration: a key enabling technology for heterogeneous co-integration of new functions with CMOS. IEEE J Emerg Sel Top Circuits Syst, 2012, 2: 714-722.

11. Hennessy J L, Patterson D A. A new golden age for computer architecture. Commun ACM,

This version posted 2026-07-03.

2019, 62: 48-60.

12. Zuo P, Lin H, Deng J, et al. Serving large language models on Huawei CloudMatrix384. arXiv preprint, 2025, arXiv:2506.12708.

13. Loh G H, Xie Y, Black B. Processor design in 3-D die-stacking technologies. IEEE Micro, 2007, 27: 31-48.

14. Chen M F, Chen F C, Chiou W C, et al. System on integrated chips (SoIC™) for 3D heterogeneous integration. In: Proceedings of 2019 IEEE 69th Electronic Components and Technology Conference (ECTC), 2019.

15. Chen Y M, Ko T, Ting K C, et al. Next generation TSMC-SoIC® platform for ultra-high bandwidth HPC application. In: Proceedings of 2024 IEEE International Electron Devices Meeting (IEDM), 2024.

16. Ingerly D B, Amin S, Aryasomayajula L, et al. Foveros: 3D integration and the use of face-to-face chip stacking for logic devices. In: Proceedings of 2019 IEEE International Electron Devices Meeting (IEDM), 2019.

17. Gomes W, Morgan S, Phelps B, et al. Meteor Lake and Arrow Lake Intel next-gen 3D client architecture platform with Foveros. In: Proceedings of 2022 IEEE Hot Chips 34 Symposium (HCS), 2022.

18. Wuu J, Agarwal R, Ciraula M, et al. 3D V-Cache: the implementation of a hybrid-bonded 64MB stacked cache for a 7nm x86-64 CPU. In: Proceedings of 2022 IEEE International Solid-State Circuits Conference (ISSCC), 2022: 428-429.

19. Vinet M, Batude P, Fenouillet-Béranger C, et al. Opportunities brought by sequential 3D CoolCube™ integration. In: Proceedings of 2016 European Solid-State Device Research Conference (ESSDERC), 2016: 226-229.

20. Santos C, Vivet P, Thuries S, et al. Thermal performance of CoolCube™ monolithic and TSV-based 3D integration processes. In: Proceedings of 2016 IEEE 3D Systems Integration Conference (3DIC), 2016.

21. Mallik A, Vandooren A, Witters L, et al. The impact of Sequential-3D integration on

semiconductor scaling roadmap. In: Proceedings of 2017 IEEE International Electron Devices Meeting (IEDM), 2017.

22. Vandooren A, Witters L, Franco J, et al. Sequential 3D: Key integration challenges and opportunities for advanced semiconductor scaling. In: Proceedings of 2018 IEEE International Conference on IC Design & Technology (ICICDT), 2018.

23. Jiang Z, Lin H, Zhong Y, et al. MegaScale: Design and deployment of a 10,000-GPU system for LLM training. In: Proceedings of 21st USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2024.

24. Narayanan D, Shoeybi M, Casper J, et al. Efficient large-scale language model training on GPU clusters using Megatron-LM. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2021: 1-15.

25. Horowitz M. Computing's energy problem (and what we can do about it). In: Proceedings of 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014: 10-14.

26. Gholami A, Yao Z, Kim S, et al. AI and memory wall. IEEE Micro, 2024, 44: 33-39.

27. Aminabadi R Y, Rajbhandari S, Zhang M, et al. DeepSpeed-inference: enabling efficient inference of transformer models at unprecedented scale. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2022.

28. Liu Y, Jiang L, Li D, et al. ByteDance Jakiro: Enabling RDMA and TCP over Virtual Private Cloud. In: Proceedings of the ACM SIGCOMM 2025 Conference, 2025: 114-128.

29. Firestone D, Putnam A, Mundkur S, et al. Azure accelerated networking: SmartNICs in the public cloud. In: Proceedings of 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2018: 51-64.

30. Liao H, Liu B Y, Chen X P, et al. UB-Mesh: a hierarchically localized nD-FullMesh datacenter network architecture. IEEE Micro, 2025, 45: 20-29.

31. Yuan Y, Peng Y, Cheung S, et al. The perspective of all-silicon photonics and systems. APL Photonics, 2025, 10: 060901.

32. Shekhar S, Bogaerts W, Chrostowski L, et al. Roadmapping the next generation of silicon photonics. Nat Commun, 2024, 15: 751.