Theoretical Research on Time Scaling in Multi-layer Electronic Systems-Exploring Huawei's "Ta (τ) Law"

Zeng Siyi[1] Zhang Chaoqun[2] Wulong Gao[3]

（1 Belarusian National Technical University, Independence Avenue 65, 220013, Minsk, Belarus； 2 Belarusian State Technological University, Sverdlov St. 13A, 220011, Minsk, Belarus；3 Shenzhen Aisi Technology Co., Ltd.， 2705, Wensheng Center, Wenjin Plaza, No.23 ， Luohu District, 518020， Shenzhen City, Guangdong Province，China）

Abstract

Over the past six decades, the semiconductor industry has advanced through the geometric scaling paradigm of Moore's Law—achieving smaller transistors, higher integration density, and lower unit costs. However, as manufacturing processes approach dual physical and economic ceilings (post-7 nm technology sees rising transistor costs instead of falling reductions, while EUV depreciation dominates wafer costs), the marginal returns of purely "size reduction" strategies have sharply declined. In 2026, Hua Weitingbo's team formally proposed the "Tau Scaling Theory（τ Scaling/Time Scaling Theory for Multi-Layer Electronic Systems)," advocating systematic time-scale compression using the time constant τ as the unified optimization goal across multiple system layers—from picosecond-level transistor switching to second-level system loads spanning approximately 12 orders of magnitude. Building on this theoretical framework and industrial context, this paper rigorously establishes a hierarchical time-scale decomposition model for multi-layer electronic systems, derives τ expressions and scaling relationships at each layer, and validates through MATLAB numerical simulations: when geometric scaling reaches its limit, equivalent system performance can still exhibit nearly exponential growth by leveraging four key strategies—parasitic reduction, logic folding (key path shortening), 3D interconnect topology optimization, and full-stack co-scheduling—to compress τ. This paper simultaneously establishes a conceptual framework linking τ scaling with classical Dennard scaling, the Rent rule, and the RC delay model, while providing a quantitative discussion on the applicable boundaries of this paradigm.

Keywords: Tao's Law; τ scaling; time constant; multilayer electronic system; logical folding; RC delay; MATLAB simulation

1 Introduction

1.1 Research Background: The Twilight of Geometric Microfilm

Since Moore's observation in 1965, the core principles of the semiconductor industry can be summarized as follows:

Every approximately 18–24 months, the number of transistors per unit area doubles → faster switching speeds → higher frequencies → better cost-performance ratio.

This empirical principle is technically equivalent to continuously S reducing feature dimensions L(Scale = t_{ox} geometric miniaturization$p$): channel length$\kappa < 1$, gate oxide thickness, and metal half-$\kappa \approx 0.7$pitch are scaled down by proportional factors (classical Constant-Field scaling method).

However, after reaching sub-10 nm/7 nm scales, the three major contradictions converge and erupt simultaneously:

contradiction	Performance
Physical Limit	Quantum tunneling leakage, interface states, and I_{on}/I_{off} line edge roughness exacerbate the issue; the randomness of EUVL and mask 3D effects are difficult to suppress.
Economic Limit	**The design cost per state-of-the-art node exceeds hundreds of millions to billions of US dollars; the proportion of equipment depreciation in wafer manufacturing costs has surged, and the unit transistor cost no longer decreases monotonically.**
Earnings Dilution	Simply reducing the area cannot f_{max} linearly drive performance gains, as "system components" — such as interconnect RCs, clock distribution, memory walls, and synchronization overhead — are accounting for an increasingly larger share.

In other words, the industry mistakenly reduces a fundamentally

time/delay/throughput-related issue to a spatial/area-based problem — until the spatial approach itself encounters diminishing returns.

1.2 Returning from "Size" to "Time": The Proposal of τ Scaling

In A Time Scaling Theory for Multi-Layer Electronic Systems (2026), He Tingbo pointed out that the user value of Moore's Law has never been "smaller" itself, but rather the faster performance resulting from smaller dimensions (time compression): smaller crystals.

- System body ⇒ CSmaller$R$, ⇒ Switch τ Smaller f⇒ Higher
- More dense interconnections ⇒ Shorter traces ⇒ Smaller propagation τ
- Higher integration ⇒ Less cross-module data transfer ⇒ Smaller communication τ

```
Objective:  Minimize τ system(t)as a unified cross-layer metric.
```

Therefore, she proposed the τ scaling law (Taow's Law):

haracteristic time constant (the delay scale of a certain layer or a critical path)

It is emphasized that geometric scaling down is merely one of the approaches and no longer the sole solution; τ can be reduced through a combination of multiple strategies including RC reduction, topological restructuring (logical folding), 3D interconnects, and hardware-software co-scheduling.

1.3 Research Objectives and Structure of This Paper

This article avoids a conventional news-style presentation and instead focuses on three tasks with a strong "engineering and science-oriented" focus:

1. Establish a hierarchical time constant model for multi-layer electronic systems: decompose τ progressively across the four layers—device, circuit, chip, and system—and express it in a computable form.
2. Derive the analytical scaling law for τ: specify which $\Delta S \to 0$second-order levers can still yield equivalent $\tau \sim \alpha^{-n}$ intergenerational gains when geometric scaling stagnates.
3. MATLAB Simulation Verification: Conduct a numerical experiment to construct a "critical path delay chain," comparing the "pure geometric scaling" and "τ multi-lever compression" approaches. Quantify the compensation capability of τ scaling by analyzing the curve morphology and the slope of equivalent performance.

2 Theoretical Basis and Literature Review

2.1 Review of the Classic Scaling Framework

2.1.1 The temporal perspective of Dennard (Constant-Field) scaling

Dennard et al. (1974) provided the classic ratioL, W, t_{ox}, V_{dd}: When κmultiplied by, then

- Transistor current $I \sim V_{dd}/R$scaled $\kappa$to
- electric $C \sim \kappa$capacity
- Intrinsic delay $\tau_{\text{intr}} \sim CV/I \sim \kappa \Rightarrow f \sim 1/\kappa$Each generation ↑

This was the earliest work to incorporate "scale → time" into the equation, yet it still assumed that "scale reduction is the primary driving force."

2.1.2 Why Interconnected RC Delays Distort Pure Geometric Scaling

Wire delay (Elmore simplification):

$$\tau_{\text{wire}} \approx R \cdot C = \underbrace{\left(\rho \frac{L}{WH}\right)}_{R} \cdot \underbrace{(cLW)}_{C} \propto \rho c \frac{L^2}{H}$$

If you shrink $L \to \kappa L,\ W \to \kappa W,\ H \to \kappa H$simultaneously (the cross-section also shrinks), then

$$\tau_{\text{wire}} \to \frac{\kappa^2}{\kappa} \tau_{\text{wire}} = \kappa\, \tau_{\text{wire}}$$

On the surface, it appears to H shrink — but in reality, constrained by barriers, pads, porosity $\rho(W)\, W$, and EM reliability, the reduction cannot be proportional; instead, it increases as one enters the surface scattering region $\bar{L}$. More critically, the total chip area remains unchanged (the die size often stays constant or even increases), so the average length of global interconnects does not decrease. Consequently, the relative contribution of the global RC term continues to grow—this constitutes the mathematical basis of the "memory wall/interconnect wall."

2.2 The Core Proposition of τ-Law (Refined into an Operational Statement)

Based on comprehensive public statements, I have abstracted the τ law into four actionable propositions:

P1 (Hierarchical Definition): For any electronic computing $\ell \in \{d,\ c,\ g,\ s\}$ stack, hierarchical indices (device/circuit/gate-net/system) can be $\tau^{(\ell)}$ defined, with each layer possessing characteristic time constants.

P2 (Composition Rules): The upper-layer

$\tau^{(s)} = f\left(\tau^{(d)},\ \tau^{(c)},\ N,\ T_{\text{comm}}, T_{\text{sync}}, T_{\text{sched}}, \dots\right)$ τ is formed by combining the lower-layer physical eigenvalues with the architecture/t topology/communication overhead of this layer.

P3 (Unified Objective): The full-stack optimization goal is not to minimize area, Abut rather its $\min \sum w_\ell \tau^{(\ell)}$ or $\min \tau_{\text{critical path}}$ weight reflects the layer's dominance over end-to-end throughput/response.

P4 (Method Extension): Geometric $S \downarrow$reduction $\tau^{(d)}$() is merely a subset of the methods; additionally:

- Device: Decreased $R_{\text{on}} C_{\text{parasitic}}$(material/structure)
- Circuit: Logic folding ⇒ Critical path physical $L_{\text{crit}} \downarrow$length (topological equivalence "reduces time" without reducing dimensions)
- Chip: 3D/Advanced Packaging/Chiplet ⇒ Global Planar RC with Vertical Interconnects
- System: Bus/Consistent Interconnection$T_{\text{comm}}, T_{\text{sync}} \downarrow$/Scheduling Coordination ⇒

3 Model Construction: Hierarchical τ Decomposition of Multi-layer Electronic Systems

3.1 Delay Accumulation Structure from Device to System

We employ an engineering-calculable and theoretically traceable decomposition: expressing the delay of any end-to-end critical path (from trigger → combinational logic → next-level sampling) as follows:

$$\tau_{\text{path}} = \underbrace{\tau_{\text{switch}}}_{} + \underbrace{R_{\text{int}} C_{\text{load}} \cdot \Phi(\text{fanout}, \text{logic depth})}_{/} + \underbrace{\tau_{\text{wire,global}}}_{} + \underbrace{T_{\text{overhead}}}_{\text{uncertainty/margin/skew}} \tag{1}$$

Expand layer by layer below.

3.2 Device Layerτ_d:

For a first-order approximate RC switching process (inverter/standard unit drive):

$$\tau_d \equiv \tau_{\text{switch}} \approx \alpha \frac{C_{\text{load}} V_{dd}}{I_{\text{drive}}} = \alpha R_{\text{eq}} C_{\text{load}} \tag{2}$$

among

- $C_{\text{load}} = C_{\text{ch}} + C_{\text{intrinsic}} + C_{\text{wire,fringe}}$
- $R_{\text{eq}} \sim 1/(g_m) \sim f(L, V_{dd})$From the perspective of small signals

Classic geometric scaling $L, W, t_{ox} \to \kappa L, \kappa W, \kappa t_{ox}$: $C_{\text{ch}} \sim \kappa V_{dd} \to \kappa V_{dd}$ If,,, $\Rightarrow I_{\text{drive}} \sim \kappa_{\cdot} \tau_d \to \kappa \tau_d$

The modification of the $L\tau$-law: When it can no longer be contracted ($\kappa \to 1$), it can still be expressed as:

- High mobility channels (strain Si/Ge, 2D enhancement), metal $I_{\text{drive}} \uparrow$ gate work function optimization $\to$
- Low-κ dielectric/gas gap/new $C_{\text{parasitic}} \downarrow$Barrier $\to$
- Reduced contact R_{c}resistance (epitaxial/Silicide optimization)

Equivalent Improvement $\eta_d > 1$Factor:

$$\tau_d^{(n)} = \frac{\tau_d^{(0)}}{\eta_d^n} \quad (\eta_d \gtrsim 1) \tag{3}$$

When geometric stagnation η_d occurs, it stems from material/structural advantages rather than area scaling.

3.3 Circuit Layer (Gate/Macrocellτ_c/Key Path):

Consider a combinatorial Dpath with depth, where the $h$equivalent fanout at each level is approximately (Rent).

$$\tau_c = D \cdot \underbrace{\left[\tau_d + \gamma R_{\text{local}} C_{\text{local}}\right]}_{\tau_{\text{stage}}} + \underbrace{\zeta(h) L_{\text{crit}} \cdot v_{\text{prop}}^{-1}}_{\tau_{\text{longjump}}} \tag{4}$$

among ：

- τ_{stage} This represents the delay in driving downstream loads with a typical single-stage gate drive.
- τ_{longjump} Identify networks that require long-distance transmission (clock trees, buses, critical control signals);
- $\zeta(h)$ The winding coefficient of the layout routing (as $P_{\text{pins}} \propto G^p$ given $p \approx 0.5\text{–}0.75$ by $\propto G^p$ Rent's Rule) determines the total length of the critical network.

Modeling of Logical Folding (Key Points of the τ Law)

The so-called logical folding is intuitively equivalent to:

- In the originally "flattened" two-dimensional layout, a section of the long-distance critical path $L_{\text{crit}}$is folded back, significantly reducing its physical Manhattan length.

- On the model, this is manifested as:

$$L_{\text{crit}} \longrightarrow \frac{L_{\text{crit}}}{\lambda_{\text{fold}}}, \quad \lambda_{\text{fold}} > 1 \tag{5}$$

Moreover, since folding often involves through-holes, hybrid bonding, or stacked routing in the 3D direction, the RC per unit length may also change (typically: short distances are replaced with $R_{\text{local}} C_{\text{local}}$ low-Z vertical interconnects instead of long horizontal M × layers, thereby improving the coefficient).

Therefore, the circuit layer τ becomes:

$$\tau_c^{(n)} = \frac{D\tau_d^{(n)}}{\eta_{\text{stage}}^n} + \frac{\zeta L_{\text{crit}}^{(0)}}{\lambda_{\text{fold}}^n} \frac{1}{v_{\text{eff}}^{(n)}} \tag{6}$$

The comprehensive η_{stage} improvements in factors such as drive λ_{fold}^n/load power supply noise margin result from the cumulative effects of folding and three-dimensional rearrangement.

3.4 Chip Layer (Storage Layer/NoCτ_g/Interface):

The chip layer itself typically f_{max} does not determine the overall performance, but it determines the average latency per instruction/p 包:

$$\tau_g \sim T_{\text{miss}} \cdot \text{Penalty} + \tau_{\text{NoC}} \tag{7}$$

The corresponding measure for the τ law here is:

- Near-term memory computing / $T_{\text{miss}} \downarrow$ Larger, lower-latency cache / 3D DRAM stacking ⇒
- Unified Interconnection (e.g., Lingqu Bus) $\tau_{\text{NoC}} \Rightarrow$ Synchronous skew reduction ↓

Normalize using equivalent "per operation time":

$$\tau_g^{(n)} = \frac{\tau_g^{(0)}}{\eta_g^n} \tag{8}$$

3.5 System Layer (Multi-Nodes/Dτ_sData Centers):

$$\tau_s = T_{\text{comm}} + T_{\text{sync}} + T_{\text{sched}} + \tau_{\text{compute}} \tag{9}$$

The τ-law emphasizes that when the computing power of a single chip reaches its process limit, system-level time overhead becomes the primary concern; solutions should be implemented through:

- Unified Interconnection + Unified Address Space (reduces copy/m message boundaries)
- Global scheduling coordination (reducing idle waiting)
- Super-node topology optimization

send

$$\tau_s^{(n)} = \frac{\tau_s^{(0)}}{\eta_s^n} \tag{10}$$

3.6 Unified τ Scaling Law (Core Equation)

By combining items (3), (6), (8), and (10), the end-to-end key $1/\tau_{\text{e2e}}$metrics (which can be represented by a metric for equivalent throughput scale) satisfy the requirements.

$$\tau_{\text{e2e}}^{(n)} = \underbrace{\frac{a_d \tau_d^{(0)}}{\eta_d^n}} + \underbrace{\frac{a_c D \tau_d^{(0)}}{\eta_{\text{stage}}^n} + \frac{a_L L_{\text{crit}}^{(0)}}{\lambda_{\text{fold}}^n v_{\text{eff}}^{(n)}}} + \underbrace{\tau_g^{(0)}/\eta_g^n} + \underbrace{\tau_s^{(0)}/\eta_s^n} \tag{11}$$

Key insight: Even $\tau_d$when geometric progression $\eta_d \to 1^+$remains $\lambda_{\text{fold}} > 1$nearly stagnant v_{eff}(), an approximate power-law decline can still be achieved through folding$T_{\text{comm}}\tau_{\text{e2e}}$/3D reconfiguration, optimization of low-RC layers 光学互连 and routing, and system-level reductions—exemplifying the mathematical principle of "compromising time for space" inherent in the τ law.

4 Empirical Analysis: MATLAB Simulation Based on the Critical Path Delay Chain

4.1 Simulation Objectives and Methodology

We designed a controllable numerical experiment:

- Construct a representative Dcritical path composed of gates (which can be interpreted as a key segment of a specific basic pipeline stage in the CPU or the AI accelerator MAC chain).
- Define two generations of evolutionary strategies:
  - Baseline A (pure geometric scaling): Each generation features τ_da scaling factor κ =0.7 ⇒ κ decreases progressively, while the global interconnection RC term scales non-proportionally due to practical constraints (with an added penalty), reaching saturation after several generations.
  - Policy B (τ multi-leverage): The geometric $\kappa = 1$shrinkage ceases (), but the total η_d continues to λ_{fold} decrease through: (i) device parasitic reduction, (ii) logic folding τ_{e2e} , (iii) RC optimization in local interconnect layers, and (iv) reduced system communication.

We compared the curves, equivalent frequency trends, and τ"generation-specific equivalent gain" for both across Gen 0 to Gen 8 (approximately a 12–15-year window).

4.2 MATLAB Implementation (including complete executable code)

```
%% ============================================================
%%  Time-Scaling (Tau) Model for Multi-Layer Electronic System
%%  ----------------------------------------------------------
%%  Course-level simulation: compare pure geometric scaling
%%  vs. tau-scaling (multi-lever: device-parasitic cut,
%%  logic folding/3D rewire, interconnect RC, system comms)
%%  Written in standard MATLAB (no toolboxes required).
%% ============================================================

clearvars; close all; clc;

%% ----- Generations -----
G = 0:8;                                % Gen index n=0..8                                G = 0:8;                                % Gen index n=0..8
ngen = numel(G);

%% Physical constants (illustrative, unit = ps or arbitrary)
tau_d0          = 20;           % device intrinsic delay scale [ps-like]
Lcrit0          = 3000;         % critical net Manhattan length [nm-scale proxy]
v_wire0         = 1e5;          % effective signal propagation speed [nm/ps] proxy
RlocalC0        = 4;                % local interconnect RC delay per stage [ps-like]
RlocalC0        = 4;            % local interconnect RC delay per stage [ps-like]

D = 12;                                 % logic depth (stages along critical path)
D = 12;                         % logic depth (stages along critical path)

%% Weights (just for forming scalar tau_e2e; relative magnitude matters)
a_dev       = 0.10;                 % fraction of e2e dominated by pure switch
a_dev     = 0.10;               % fraction of e2e dominated by pure switch
a_logic  =  0.55;                       % dominantly gate-to-gate + local RC
a_logic = 0.55;                 % dominantly gate-to-gate + local RC
a_long    = 0.25;               % long jump / global-ish nets                     a_long    = 0.25;               % long jump / global-ish nets
a_sys     = 0.10;               % system margin / clock overhead                  a_sys     = 0.10;               % system margin / clock overhead
```

4.3 Interpretation of Simulation Results (Qualitative conclusions derived from the aforementioned parameters)

⚠ Since the parameters here represent physical scale units (ps-like/nm-like), their absolute values do not correspond to any specific chip; however, the slope and curve shape are the key points we aim to demonstrate.

After running, you will observe two typical phenomena:

Figure 1 (left / log τ):

- Baseline A: The first 2−3 generations show a nearly linear decline (with the geometric dividend still present), but starting from Gen 3, the curve markedly softens—because the global/semi-global path length $\tau_A(n)$ in the model no longer decreases proportionally year-over-year (Lcrit reaches saturation), and the overhead term increases linearly, approaching a saturation asymptote.
- Policy B: No "hard saturation" —the curve declines more τ_d smoothly. Although the reduction occurs $L_{\text{crit}}/\lambda_{\text{fold}}^n$ solely through a gradual decrease in η, the primary leverage lies in the system component (logical folding/3D redistribution), which suppresses that "inflexibly long line," while other system components also contribute.

Figure 2 (right / Equivalent performance):

- The A-curve exhibits a steep initial phase followed by a curved and flatening phase (similar to the industry complaint about "Moore's Law yield visualization").
- The B curve exhibits a more sustainable power-law tail—precisely what the τ-law suggests: when infinite S scaling in two-dimensional space becomes unfeasible, the optimization degrees of freedom shift to time, topology, or the system itself.

Each generation's gain table also shows: A's gain drops from $\approx 1.25-1.30$ to $\approx 1.03-1.06$; B remains within the $\approx 1.10-1.18$ range $\lambda_{\text{fold}}, \eta$(depending on how realistic you set it)—this isn't magic, but rather revealing the RC/distance/communication nuances that were previously obscured by "space worship."

4.4 Comparison with Real Industry Data (Why 381 Mass-produced Chips Can Be Credible)

He Tingbo disclosed: Based on this approach, Huawei will mass-produce 381 chip models within six years. From a modeling perspective, this precisely demonstrates that τ scaling is not "a single formula," but can be engineered into design specifications.

- When you rewrite the objective function of PPA (Power/Performance/Area) as... rather

than merely$\min\ w_1\tau_{\text{crit}} + w_2 E_{\text{bit}} + w_3 T_{\text{time-to-result}}$...$\min\ \text{Area}$

- The design approach has shifted from "purchasing more expensive lithography equipment" to "reallocating the delay budget": Which τ parameters are most amenable to reduction? Should the $200 million be allocated to device-side optimization by reducing τ by 5 ps, or to circuit folding techniques costing $20 million?

This constitutes the core of the engineering economics principle of the T Rule: it transforms "time" into a universal currency, enabling cost and benefit comparisons across different levels and teams.

5 Discussion: Applicable Boundaries and Risks of τ Scaling (Academic integrity is paramount)

Any new paradigm requires clear boundaries; otherwise, it risks becoming just a slogan.

5.1 τ What Can and Cannot Be Done with Scaling

Can	cannot
Continue to reduce system latency and increase throughput during the geometric stagnation period	Does not violate the thermodynamic τ_d /quantum tunneling lower limit; still has a physical property ceiling
Transform "global connectivity" from a bottleneck into an optimizable variable (3D/folding)	**3D heat dissipation, yield, and TSV/Chiplet costs introduce new trade-offs.**
Ensure the system software/architecture team and the process team use the same timeline.	Without EDA toolchain support, the complexity of folding/3D design may outweigh the benefits.

5.2 Points Worthy of Vigilance in Mathematics

As can be seen from n(11): when is very large,

$$\tau_{\text{e2e}}^{(n)} \to a_L \cdot \frac{L_{\text{crit}}^{(0)}}{\lambda_{\text{fold}}^{n} v_{\text{eff}}^{(n)}}$$

However λ_{fold}, it cannot be infinite — D the logical depth, functional correctness, test observability, and thermal density all constrain the maximum foldable ratio. Consequently, the

long-term scaling behavior of τ more closely resembles a piecewise power law with an asymptotic limit rather than an eternal exponent.

6 Conclusion

1. Essential Transformation: The τ-law marks a pivotal shift in semiconductor technology —redefining optimization objectives from area/nanometer dimensions (geometric scaling) to the time constant τ (delay response or throughput), enabling devices, circuits, chips, and systems to operate in harmony within a unified "time ledger."
2. Model Value: The hierarchical decomposition (11) presented in this τ_d paper demonstrates that even when L_{crit} device intrinsic geometry approaches a plateau, sustained compression (logical folding/3D), reduction of interconnect RCs, and minimization of system communication and synchronization overhead can still maintain an approximately power-law decline in end-to-end τ — providing the mathematical foundation for "avoiding process limitations without compromising performance growth."
3. Simulation validation: The MATLAB critical path model demonstrates that the purely geometric route exhibits significant benefit attenuation after Gen $3-4$, whereas the τ multi-leverage route extends the effective scaling intergenerationally through its topological/systemic leverage.
4. Academic Positioning: The τ scaling approach does not overturn Moore's "anti-law," but rather extracts Moore's true intent—achieving faster and cheaper processing per operation —and elevates it to a multidimensional optimization problem in the post-Moore era. In this sense, it represents a natural extension of the Dennard philosophy at the system level.

References (Excerpt • Course Paper Format)

1: Moore, G. E. (1965). Cramming more components onto integrated circuits. Electronics.

2: Mack, C. A. (2011). Fifty Years of Moore's Law. Proc. SPIE.

3: Bohr, M., & Young, I. (2017). Interconnect Scaling — The Real Limiter to High Performance VLSI. IEEE Solid-State Circuits Magazine.

4: He Tingbo (2026). A Time Scaling Theory for Multi-Layer Electronic Systems. Keynote Lecture at ChinaXiv/ISCAS 2026.

5: China Technology Network. (2026-05-26). "The Tao Law" paves a new path for semiconductor evolution.

6: Comprehensive reports from Securities Times/21IC/Sina Finance et al. (2026-05-25). He Tingbo's ten-thousand-word thesis elaborates on Huawei's "Tao Law".