A Time Scaling Theory for Multi-Layer Electronic Systems

Authors: Tingbo He

Date: 2026-05-25T09:56:20+00:00

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% power-efficiency gain 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.

Full Text

Preamble

A Time Scaling Theory for Multi-Layer Electronic Systems Tingbo He Huawei

Abstract

For six decades, Moore' s geometric scaling drove progress in semiconductors. That industry compact no longer holds: returns from pure dimensional shrinking have attened, 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 —scaling that adopts time itself, rather than transistor area, as the primary metric of progress, applying a single characteristic time constant 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% power-e ciency gain at a xed device node. On AI systems, a co-designed stack comprising the memory-semantic Uni ed Bus fabric, near-packaged Hi-ONE optical I/O, and edge-to-surface 3D Folding projects more than 100× growth in hardware scaling is the rst scaling principle since Dennard to establish a shared optimization target across the entire computing stack.

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 attened —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 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 scaling is developed below as both a scienti c methodology and an industrial roadmap, drawing on lessons from 381 chips brought to volume production between May 2020 and May 2026. 1. 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 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 Together, geometric scaling and Dennard scaling delivered exponential improvements in

performance per watt and performance per dollar for nearly ve decades.

This arrangement unraveled in two stages. Around 2005, Dennard scaling broke rst: 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 attened. 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.

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 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. 2. Time, Not Space: The Real Currency of Moore' s Era Reduced to its essential e ect 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 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 can be de ned 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 rather

than the only This principle is called scaling , and is proposed here as the successor to geometric Moore scaling as the guiding principle of semiconductor evolution. Formally, is treated as a layered construct that decomposes as where , and represent the time constants at the transistor, circuit, chip, and system layer, respectively.

Each layer' s composed from the layers beneath it together with the organizational and communication overheads introduced at that layer. The working space of 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- 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.

Circuit: RC propagation delay along signal paths, addressed through lower-resistivity conductors, low- dielectrics, and —most consequentially —through reduction of wire length via vertical integration.

Chip: compute and memory-access latency, addressed through architectural choices, pipeline depth, memory hierarchy, and on-chip fabrics.

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

A useful generational rule emerges from this layered formulation:

$$+1 = \quad \alpha\alpha$$

where the scaling factor application-specific rather than universal. Production experience to date indicates α 1.3× per year for power- constrained mobile devices, 1.5× per year for safety critical autonomous systems, and up to 10× per year for AI workloads , where throughput translates directly into economic value.

What renders a useful primary metric, rather than a relabeling of existing ones, is that it is same metric across the entire stack . Frequency, latency, bandwidth, and throughput are all governed at their respective layers. A process technologist, a circuit designer, and a system architect can debate the same quantity in identical units. 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.

3. LogicFolding

: A Mobile-SoC Proof Point rst production-scale test of 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 xed, 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.

Digital circuits divide into combinational logic —the Boolean network between registers —and sequential logic —the ops that hold state. The performance ceiling of a digital system is set by the critical-path delay between adjacent op 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 eventually more) vertically stacked active tiers , connected through ultra- ne-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.

To help LogicFolding deliver these gains, it is advantageous to keep the gear ratio between hybrid- bonding pitch and top-metal pitch comparatively low — roughly below 3 in practice, with lower ratios generally better . With today' s top-metal pitch around 720 nm, this translates into a hybrid- bonding pitch below 2 —and ideally to a gear ratio of approximately 1, at which the bird-cage routing overhead at the bonding interface e ectively vanishes. Achieving this pitch, together with the required overlay accuracy (<0.5 m), TSV scaling (CD and KOZ sub-1.5 m, pitchsub-6 and yield (~100% with smart redundancy), required a multi-year process-development e ort across the supplier and partner ecosystem.

The results, measured on Kirin 2026, are concrete:

Transistor density rose step-wise from 155 to 238 MTr/mm^2 in a single generation (transistor density is calculated using the formula ∗ ; the area utilization of Kirin SoC design is 68%) —a magnitude of improvement that previously required three years of geometric scaling.

SoC performance-core power efficiency improved by 41% maximum clock frequency

rose by nearly 13% A high-speed global Network-on-Chip data path constructed across both upper and lower tiers reduced the data-path footprint by , with improved power-delivery stability.

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- er count by more than 50% , clock skew by , and wire length by approximately 30% 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.

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

3.1 GHz

this year. 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 (Table 1) . The roadmap is feasible and, in cost terms, economically viable.

Trend of the operating frequency of Kirin CPU performance core.

Sidebar A —LogicFolding at a Glance Hybrid-bonding pitch: sub-2 m (1.5 μm in Kirin 2026; target gear ratio 1) Overlay accuracy: under 0.5 TSV CD/KOZ: sub-1.5 m; pitch sub-6 m; failure rate <100 ppm; repair rate 99.9% Yield: ~100% with smart redundancy Transistor density: 155 238 MTr/mm^2 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% 4. From Picoseconds to Microseconds:

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 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 a rmative —provided is treated as a system-level objective and applied across the whole chain, rather than within a single accelerator.

Two facts shape the AI side of the argument. First, AI systems continue to grow —from one chip, to dozens, to hundreds, and increasingly to tens of thousands.

Second, the energy budget and the materials budget of modern AI systems are dominated by data, not by compute.

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 scaling is instantiated at AI scale through three coordinated layers: a system fabric (Uni ed Bus), a near-packaged optical engine (Hi-ONE), and a topological reorganization of the package itself (3D Folding).

4.1 Uni

fi ed 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 In niBand between chassis, and software-stack remote-memory access on top. Each layer entails a protocol conversion, additional serialization, an extra DMA bu er, 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 across the chassis —a fully peer-to-peer fabric that exposes memory semantics natively across the whole system. 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 bene t 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 ~500× reduction in system along the dominant communication axis. At the rack scale, this brings the system asymptotically close to a single, fabric-coherent machine —designated internally as a System-as-One-Chip

4.2 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- delivery margins are exhausted.

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 . 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 -scaling argument. In place of a heavy DSP for high signal delity, Hi-ONE adopts a linear approach —an analog equalization-enhanced driver and trans-impedance ampli er —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-o that a methodology rewards. 4.3 The N^2-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 compute resources within.

If the die has side length N, then: compute capacity scales as (area), but memory bandwidth, interconnect, and power delivery —all carried by the 2.5D fan-out along the edge —scale only as (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 , 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 through 2035. Along this path, hardware integration is projected to increase by than100× by 2035 , with reduction distributed across every layer of the stack rather than concentrated at the device level.

Sidebar B —at AI System Scale UB remote-access latency: ~10s of μs $\rightarrow$ ~100 ns (500× τ reduction) HiONE per-module bandwidth: 8 Tb/s (matches per-chip UB bandwidth) HiONE SerDes reach: ~100 cm $\rightarrow$ ~5 cm; panel panel reach: <1 m $\rightarrow$ 100 m Fan-out dilemma: compute N^2, perimeter-bound BW/I/O/power 3D Folding: relocates BW, optical I/O, and power delivery from edges onto surfaces, restoring N^2 parity

- 2026 $\rightarrow$ 2035 projected hardware -integration growth: >100×

5. Logic and Memory: From Decoupling to Re-Fusion One implication of 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 in uence 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 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, scaling ensures that the problem cannot be deferred.

6. Open Challenges

It would be misleading to present scaling as a completed system. Several substantive problems remain open, and are identi ed 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 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 uni ed 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. -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 , 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 -aware signo ff fl ows are necessary components of the response.

Vertical-interconnect overhead. Every hybrid bond and every TSV incurs a nite resistance and capacitance penalty, and TSV KOZ displaces standard cells.

LogicFolding must therefore be justi ed layer by layer through the simple inequality 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. is a time law, not a joule law. A super-node operating 10× faster but with 10× greater power consumption violates no scaling principle, yet exceeds grid capacity. 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 headroom back for power

(DVFS at data-center scale —the same mechanism that enabled smartphone

battery longevity).

Importantly, 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 su -scaling industry requires profile benchmarks —vectors that expose the dominant at each layer of a system together with the headroom remaining at that layer. The dominant-layer is, by de nition, the next investment. 7. 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 scaling thesis has held up:

At the device and circuit layers, transistor density has risen from 155 toward 400+MTr/mm^2 by 2031.

At the chip layer, LogicFolding has demonstrated, on a leading-edge mobile SoC, that critical- path frequency, power e ciency, and density can continue to advance at a xed device node.

At the system layer, ed Bus and Hi-ONE have demonstrated that hundreds of microseconds of communication 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 e ciency is projected to more than double in three to ve 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. scaling is the first scaling 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 to count. It also indicates to industry strategists and

capital allocators that the next dollar should follow 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 di cult 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 optimization across the multi-layered electronic system —and the companies, research groups, and ecosystems that adopt as the primary ob-

jective 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.

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 scaling methodology and the LogicFolding

Uni fi edBus, and Hi-ONE technologies described in this article.

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.

Further Reading 1. G. E. Moore, "Cramming more components onto integrated circuits," Electronics , vol. 38, no. 8,pp. 114–117, Apr. 1965 (reprinted in Proc. IEEE , vol. 86, no. 1, Jan. 1998).

2. R. H. Dennard

et al. , "Design of ion-implanted MOSFETs with very small physical dimensions," IEEE J. Solid-State Circuits , vol. 9, no. 5, pp. 256–268, 1974. 3. J. L. Hennessy and D. A. Patterson, "A new golden age for computer architecture," Commun. , vol. 62, no. 2, pp. 48–60, Feb. 2019. 4. M. Horowitz, "Computing' s energy problem (and what we can do about it)," ISSCC Dig. Tech.

Papers , pp. 10–14, Feb. 2014. 5. International Roadmap for Devices and Systems (IRDS) —Interconnect and More-than- Moore chapters, 2023/2024 update.

6. P. Batude

et al. , "3D sequential integration: a key enabling technology for heterogeneous co- integration of new functions with CMOS," IEEE J. Electron Devices Soc. , vol. 3, no. 3, pp. 205–216, 2015.

Note: Figure translations are in progress. See original paper for figures.