Physical Dilemma of Large-Area Advanced-Node Chips: Irreversible Narrowing of Interconnect Channels

Ao Li

Independent Researcher

Author Biography

I am a high school student dedicated to researching chip development paths in the post-Moore's Law era. I have developed a complete, implementable solution based on Mature Nodees and chiplets. From geometric constraints and fundamental physical limits, I have systematically proven the inherent bottlenecks of advanced-node scaling. Due to limited funding for patent applications, full technical details are temporarily undisclosed to protect core intellectual property. I welcome inquiries for potential collaboration.

Email: hkts_88@qq.com

Abstract

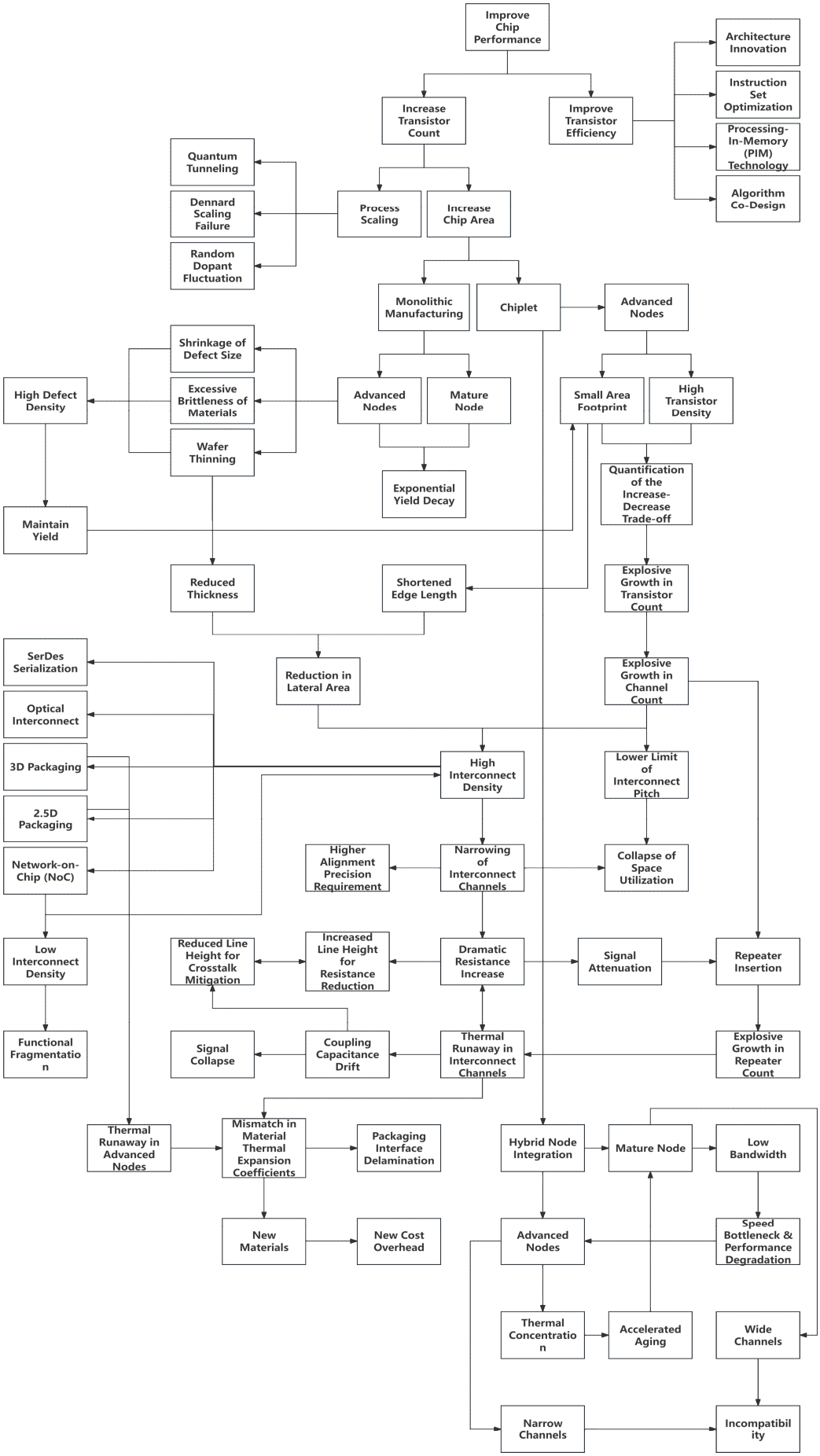
The two pathways to increasing chip computing power—process scaling and chip area expansion—have both reached their physical and economic limits. This paper demonstrates that the physical bottlenecks of advanced nodes and the contradictions of interconnect narrowing cannot be resolved by 2.5D/3D packaging or chiplet architectures, which merely shift costs rather than eliminate fundamental physical constraints. Industrial remedies are shown to create markets out of their own self-inflicted problems. As sub-1 nm nodes approach, uncontrollability overtakes usability. The only physically self-consistent path forward is a system-integration paradigm centered on mature-node chiplets, where wide interconnects, low thermal density, and traditional packaging together respect fundamental physical laws.

Keywords

Advanced Node; Interconnect Channel; Narrowing Effect; Resistance; Thermal Dissipation; Chiplet; Mature Node; Physical Limit; RC Delay; Controllability

---

Mind Map


Improve Chip Performance
Architecture Innovation
Instruction Set Optimization
Processing-In-Memory (PIM) Technology
Algorithm Co-Design
Increase Transistor Count
Improve Transistor Efficiency
Quantum Tunneling
Dennard Scaling Failure
Random Dopant Fluctuation
Process Scaling
Increase Chip Area
Monolithic Manufacturing
Chiplet
Advanced Nodes
Shrinkage of Defect Size
Excessive Brittleness of Materials
Wafer Thinning
High Defect Density
Advanced Nodes
Mature Node
Small Area Footprint
High Transistor Density
Exponential Yield Decay
Quantification of the Increase-Decrease Trade-off
Maintain Yield
Reduced Thickness
Shortened Edge Length
Explosive Growth in Transistor Count
SerDes Serialization
Optical Interconnect
3D Packaging
2.5D Packaging
Network-on-Chip (NoC)
Reduction in Lateral Area
Explosive Growth in Channel Count
High Interconnect Density
Lower Limit of Interconnect Pitch
Higher Alignment Precision Requirement
Narrowing of Interconnect Channels
Collapse of Space Utilization
Low Interconnect Density
Reduced Line Height for Crosstalk Mitigation
Increased Line Height for Resistance Reduction
Dramatic Resistance Increase
Signal Attenuation
Repeater Insertion
Functional Fragmentation
Signal Collapse
Coupling Capacitance Drift
Thermal Runaway in Interconnect Channels
Explosive Growth in Repeater Count
Thermal Runaway in Advanced Nodes
Mismatch in Material Thermal Expansion Coefficients
Packaging Interface Delamination
Hybrid Node Integration
Mature Node
Low Bandwidth
New Materials
New Cost Overhead
Advanced Nodes
Speed Bottleneck & Performance Degradation
Thermal Concentration
Accelerated Aging
Wide Channels
Narrow Channels
Incompatibility

0 The Only Two Paths to Performance Improvement: A Geometric Fact

Before any technical discussion, it is necessary to establish a premise dictated by geometry and basic physics: under current physical laws, there are only two fundamental paths to increasing chip computing power.

Path 1: Process scaling—increasing transistor density per unit area. Accommodating more transistors within the same chip area to handle more simultaneous computation. This is the core logic of Moore's Law: continuously shrinking transistor feature size to integrate more computing units into the same silicon piece [1].

Path 2: Chip area expansion. Keeping transistor density constant and simply manufacturing a larger chip, trading total area for total transistor count.

There is no third path. The paths referred to here are specifically the means of transistor integration at the physical hardware level. Architectural innovations, instruction set optimizations, processing-in-memory, algorithmic co-design, and similar approaches improve the computational efficiency per transistor; they do not alter the

upper bound on the total number of transistors a chip can accommodate, and therefore do not break the two-dimensional geometric constraint [1].

This conclusion is rigorously derived from geometry: a chip is a two-dimensional physical entity, and its total transistor count equals "area" multiplied by "transistor density per unit area." Any so-called "architectural innovation" or "heterogeneous integration" is ultimately a combination or compromise of these two paths and cannot escape this planar geometric constraint. Forcibly stacking chips vertically into a three-dimensional structure does not fundamentally break the area-perimeter geometric lock; instead, it worsens the heat extraction path from a two-dimensional planar problem into a three-dimensional heat conduction dead end [2].

The following analysis will demonstrate: under advanced nodes, Path 1 encounters quantum tunneling [3], the Dennard scaling limit [4], and uncontrolled statistical fluctuations; the monolithic large-chip version of Path 2 encounters an exponential yield degradation bottleneck [5][6]. Any attempt to forcibly splice the two paths through 2.5D packaging or chiplet architectures can shift some contradictions, but it introduces more complex systemic challenges at new physical interfaces [2]. The key question is whether the industry has the courage to admit that advanced

nodes are not the default optimum for all scenarios.

---

1 The Physical Bottlenecks of Process Scaling

For the past half-century, the improvement of chip performance in the semiconductor industry has been guided jointly by Moore's Law and Dennard scaling. However, as process scaling has caused Dennard scaling to fail first, further scaling now faces the following predicaments.

1.1 Quantum Tunneling

When transistor channel length approaches the atomic scale, electrons can tunnel through potential barriers with significant probability even in the off-state [3]. The switch can no longer be fully turned off, and leakage current increases exponentially with size reduction. This is not a material defect; it is a fundamental prediction of quantum mechanics.

1.2 Dennard Scaling Limit

Dennard scaling once allowed voltage to be reduced proportionally with dimensions, maintaining constant power density as transistor density doubled [4]. However, the threshold voltage cannot be lowered infinitely due to thermodynamic limits—below a critical value, off-state leakage drowns the switching signal. When voltage stops scaling while density

continues to double, power density doubles. This is a red line drawn jointly by Joule's law and thermodynamics.

1.3 Uncontrolled Statistical Fluctuations

When the number of dopant atoms in the channel drops to the order of a few dozen, random fluctuations cause each transistor's threshold voltage to differ. The slowest device on the chip drags down the overall frequency, while the fastest accelerates aging due to over-driving. Yield and reliability, at the atomic scale, become victims of statistics.

Three independent physical laws—quantum mechanics, thermodynamics, and statistical physics—each indicate that process scaling as a single path has reached its physical limits of applicability.

---

2 The Yield Predicament of Advanced Nodes

In semiconductor manufacturing, chip yield follows the Poisson yield model:$Y = e^{-DA}$(where D is the defect density and A is the chip area) [5]. However, advanced nodes inherently face a higher defect density in the manufacturing process than mature nodes. The following factors make the actual yield situation far more pessimistic than the theoretical prediction of the Poisson model.

2.1 Dicing Damage Induced by Wafer Thinning

To alleviate the thermal pressure caused by high power density, chips are often forced to be thinned. However, the thinned wafer is more prone to edge chipping during dicing, and the microscopic particles generated, once landing on the chip surface, can form fatal defects.

2.2 Mechanical Fragility of Low-k Dielectric Materials

To reduce RC delay, advanced nodes widely adopt low dielectric constant (low-k) materials. These materials are porous in texture and poor in mechanical strength, making them highly susceptible to dielectric layer delamination during dicing, further driving up the defect density.

2.3 Synchronous Shrinking of the Critical Defect Size

As the process scales down, the critical size of a fatal defect shrinks synchronously. Microscopic particles that could previously be ignored are now sufficient to scrap an entire chip [5]. The defect density at the wafer edge is inherently higher than that in the central region. Large-area chips, due to their larger area, have a higher probability of covering these edge regions, making the actual yield situation even worse. Consequently, the

industry is compelled to reduce chip area to maintain yield.

2.4 The Hidden Cost of Yield Improvement

In the face of the above predicaments, the industry is not standing idly by. To pull the yield of advanced nodes back into an economically viable range, foundries continuously insert more defect inspection, redundancy repair, and process compensation steps into the manufacturing flow.

However, these additional process steps are not a free lunch. Each additional step drives up manufacturing costs, causing wafer quotations to climb steadily [6]. What is worse, as the process shrinks to the atomic scale, the physical root causes of defects can no longer be eradicated by process optimization and can only be managed statistically.

While more inspection steps improve yield, they also introduce new inspection errors and process variations. The yield improvement curve tends to flatten, with diminishing marginal returns. Yield management for advanced nodes is degrading from "eradicating defects" into "trading cost for yield"—using ever-increasing manufacturing costs to exchange for ever-weaker yield improvements [5][6].

The above analysis collectively points to a core conclusion: under advanced nodes, directly manufacturing large-area monolithic chips is physically and economically infeasible. Even if foundries invest in yield management regardless of cost, once the single-die area exceeds a certain threshold, the yield will collapse to a commercially unacceptable level. This means the traditional path of expanding single-chip area has been thoroughly blocked by the laws of physics.

---

3 The Contradiction Between Chiplets and Advanced Nodes

Facing the yield problem of monolithic chips, the industry has collectively turned to the chiplet approach: splitting the functionality of a large chip into multiple small dies and reassembling them into a complete system through advanced packaging [7]. Although this route circumvents the yield trap of monolithic chips, it transfers the interconnect contradictions previously hidden inside the chip to the packaging interface.

An increase in transistor count means that the total number of signals that need to be processed inside the chip increases correspondingly. This relationship is not an arbitrary assumption, but a fundamental law described by Rent's rule: the external interconnect demand of a system

follows a power-law relationship with its internal complexity:$T = k \cdot N^P$ (where *T* is the number of external interconnect channels, *N* is the number of internal modules, and *P* is the Rent exponent). In logic chips without specialized architectural optimization, *P* approaches 1, and the external interconnect demand grows approximately linearly with the internal scale.

This is precisely the real predicament facing advanced nodes, and there is no middle ground to escape it. The original purpose of chiplets was to resolve the yield predicament of monolithic large chips—splitting a large chip into multiple small dielets and reassembling them into a logical whole through advanced packaging. However, the premise of this "reassembly" is that sufficiently dense interconnect channels must be preserved between dielets so that the split system can logically approximate a single complete chip. In other words, the ideal interconnect state for the chiplet approach is precisely *P* approaching 1—global interconnect, arbitrary communication, and intact logical wholeness.

Yet, if one attempts to suppress the value of *P* through architectural optimization in order to reduce the number of interconnect channels, one immediately falls into another trap: a suppressed *P* means that the majority of signals are confined inside each dielet, and each dielet

degenerates into an isolated functional island. A system assembled from multiple dielets no longer logically approximates a single complete large chip, and the act of reassembly itself loses its meaning.

As for compressing *P* to some intermediate value—retaining a portion of global interconnect to maintain fragile wholeness, while compressing a portion of channels to avert physical disaster—this merely incurs the costs of both sides simultaneously: paying the expensive cost of advanced packaging while failing to achieve the logical integration density of a monolithic chip.

The industry wants both the high transistor integration density of advanced nodes and the ample area of a large chip, yet these two objectives are mutually exclusive under the physical constraints of interconnect narrowing. This is the trilemma built into the chiplet approach from the day of its inception.

When *P* approaches 1, the system has lost effective modular hierarchy internally, and every basic logic unit directly requests communication with the outside. At this point, the "number of internal modules" *N* in Rent's rule no longer refers to high-level functional modules, but has degenerated into the number of logic units at the lowest level—whose

growth is approximately linear with the transistor count.

In other words, in the limiting state of P approaching 1, every doubling of transistor count is accompanied by a near-proportional escalation in equivalent external communication demand. This is the physical reality that advanced-node chips cannot circumvent: density increases→ transistor count rises→equivalent module count increases→external interconnect demand explodes synchronously. Therefore, P approaching 1 is the logically ideal state for the chiplet approach, and it is also the reality it must confront. The following quantitative analysis is conducted based on this state.

To preserve yield and facilitate heat dissipation, the area and thickness of advanced-node chips are compressed in both directions: the shrinking area reduces the side length, and the thinning reduces the sidewall height —while the number of channels is exploding, the lateral surface area available for arranging these channels is drastically shrinking. To quantify the extent of this "double squeeze," the lateral surface area of the chip must be brought into consideration. Since interconnect channels are primarily distributed on the sides of the chip, the lateral surface area directly determines the upper limit of the physical space for pin layout. The lateral surface area equals the perimeter multiplied by the

thickness:$S = C \times H = 4\sqrt{A} \times H$(where C is the perimeter, A is the chip area, and H is the chip thickness).

Substitute real product data for comparison. The following is the CCD chip information for AMD under a mature node (28 nm) and an advanced node (5 nm):

Item	Mature Node(28 nm)	Advanced Node(5 nm)
Chip Example	AMD Carrizo APU	AMD Zen 4 CCD
Chip Area (A)	ca.245 mm^2	ca.70 mm^2
Transistor Count (N_t)	3.1 billion	6.57 billion
Chip Thickness (H)	ca.250 μm	ca.75 μm

Calculation Process

① Side Length and Lateral Surface Area

Mature node (28 nm):

$$S = 4\sqrt{A} \times H = 4\sqrt{245} \times 0.25 \approx 15.65\, mm^2$$

Advanced node (5 nm):

$$S = 4\sqrt{A} \times H = 4\sqrt{70} \times 0.075 \approx 2.51\, mm^2$$

② Transistors per Unit Lateral Surface Area

Mature node:

$$\frac{N_t}{S} = \frac{3.1\,billion}{15.65\,mm^2} \approx 0.198\,billion/mm^2$$

Advanced node:

$$\frac{N_t}{S} = \frac{6.57\,billion}{2.51\,mm^2} \approx 2.62\,billion/mm^2$$

③ Ratio

$$\frac{(N_t/S)_{adv}}{(N_t/S)_{mat}} = \frac{2.62}{0.198} \approx 13.2$$

The calculation result reveals the core of the problem: the number of transistors that a unit lateral surface area of an advanced-node chip needs to accommodate is approximately 13 times that of a mature-node chip.

This is a very intuitive physical squeeze. Although the chip area has shrunk by 71% (from 245 mm^2 to 70 mm^2), the transistor count has doubled (from 3.1 billion to 6.57 billion). To control power consumption and facilitate heat dissipation, the chip has been further thinned (thickness reduced from about 250 μm to about 75 μm).

This series of operations has produced two fatal consequences. On one hand, the increase in transistors demands more chiplet I/O channels. On the other hand, the thinning and narrowing of the chip have caused the lateral surface area used for arranging I/O pins to be extremely compressed. Under this trade-off, the interconnect pressure increases by

an order of magnitude, making the congestion of interconnect channels an unavoidable physical necessity.

This result directly dismantles the narrative foundation of the chiplet approach. The industry had hoped that chiplets could "split and dissolve" the contradictions, but the laws of physics coldly indicate that the interconnect pressure has not been split and diluted; rather, it has been further concentrated and intensified in the microscopic world of the chip. The advantage of advanced-node chips in transistor density ultimately translates into a several-fold to dozens-of-fold higher interconnect density requirement at the packaging interface [7].

---

4 Limitations of Industrial Remedies

Section 3 has already fully deduced the logical consequences of lowering the value of P: when P approaches 0, dielets degenerate into functional islands, and the "wholeness" of the chiplet approach ceases to exist. This section does not repeat that argument, but instead examines, one by one, four technical approaches that the industry has attempted to use to break the relationship between transistor count and channel count. Among them, some approaches attempt to reduce the value of P through architectural means, while others do not change P but trade other costs for relief in channel count. However, upon closer examination, each approach merely

uses new costs to shift old contradictions, without eliminating the problem of interconnect narrowing at its root.

4.1 SerDes Serialization—Trading Space for Time

SerDes serialization does not change the value of *P* in Rent's rule. What it does is not reduce the demand for global interconnect, but combine multiple low-speed parallel signals into a single high-speed serial signal, and then de-serialize them at the receiving end. Data that originally required 16 channels can be reduced to just 2 to 4 channels after serialization.

This is a classic case of "trading space for time"—the physical channel count is indeed reduced, but the total signal volume remains unchanged, and the total bandwidth requirement remains unchanged. The cost is extremely high power consumption in the serial link itself (the power consumption of high-speed SerDes can account for a considerable proportion of a chip's total power budget), along with a dramatic increase in signal integrity design complexity. It mitigates the explosion in channel count, but it does not sever the relationship between transistor count and total signal volume at its root.

4.2 Network-on-Chip and Localized Computing—Cutting the Necessity for Signals to Leave the Chip

This is currently the industry's primary means of suppressing the value of P : through Network-on-Chip (NoC) and distributed computing architectures, the majority of signals are digested inside the chip, with only the final results needing to be communicated externally. This directly reduces the total number of signals that must leave the chip at the source, severing the relationship between transistor count and channel count. A more radical direction is Processing-in-Memory—moving computation to where the data resides, completely eliminating the need for data transport.

However, as Section 3 has already analyzed, suppressing the value of P inevitably leads to functional islands. When the majority of signals do not leave the dielets, a system assembled from multiple dielets is in essence no different from a group of independent chips, and the "logical wholeness" sought by the chiplet approach ceases to exist. This is not a difficulty that can be circumvented through engineering optimization, but a logical consequence inherent in the very act of lowering the value of P .

4.3 3D Integration—Breaking the Planar Bottleneck with the Vertical Dimension

3D stacking does not change the value of *P*, but attempts to physically increase the channel supply. Traditional chips can only arrange I/O along the planar perimeter, whereas 3D stacking can place interconnect channels across the entire surface of the chip face, with interconnect densities far exceeding the upper limit of planar arrangements. This is not about reducing channel demand, but about opening up an entirely new physical dimension to carry interconnects.

The costs have already been thoroughly analyzed in Section 10: thermal dead-ends and structural warpage. Heat from middle layers cannot escape, and thermal expansion mismatch tears apart microbumps—this is a physical deadlock that 3D packaging cannot fundamentally cure.

4.4 On-Chip Optical Interconnects—The Ultimate Solution of the Future?

Silicon photonics technology attempts to replace copper wires with optical waveguides for signal transmission. The bandwidth density of optical interconnects far exceeds that of copper wires, and the transmission power consumption is almost independent of distance. If

future inter-chip interconnects switch from "electrical" to "optical," the physical bottleneck of channel count would be fundamentally circumvented. However, this technology is still at the laboratory stage and is a considerable distance from mass production.

The problem is that all of the above methods add additional costs, which are ultimately borne by consumers. The problems themselves, however, are caused by advanced nodes—the industry is using problems of its own making to create new markets, which has already extended well beyond the scope of technical discussion.

---

5 Channel Narrowing Causes Drastic Resistance Increase

The quantitative analysis in Section 3 has already shown that the number of transistors per unit lateral surface area in an advanced-node chip is approximately 13 times that of a mature-node chip. The lateral surface area is the upper limit of the physical space available for I/O pin layout—the same lateral surface area must now serve 13 times as many transistors, meaning that the number of interconnect channels that must be accommodated per unit lateral surface area has also exploded by a factor of approximately 13. The cross-sectional area of each channel is thus forcibly reduced to approximately 1/13 of its original value. According to

the resistance law $R = \rho L/(w \cdot \tau)$ (where w is the interconnect line width and τ is the interconnect line height), the resistance of a channel is inversely proportional to its cross-sectional area—when the cross-sectional area shrinks to 1/13, the resistance directly increases by a factor of 13.

In reality, the situation is worse than this estimate suggests. The surface of a conductor cannot be absolutely smooth. When the line width shrinks to the nanometer scale, and the degree of surface smoothness is subject to an upper limit imposed by process constraints, electrons moving within the conductor inevitably and frequently collide with the rough sidewalls. Each collision causes a loss of momentum, equivalent to an additional increase in resistance [8][9].

At the same time, the interior of a copper conductor is not a single perfect crystal, but is instead composed of countless small grains. The boundaries between these grains also scatter electrons. The narrower the line width, the smaller the grain size and the higher the grain boundary density, which means a greater probability of electrons colliding with grain boundaries. These two scattering effects superimpose, causing the effective resistivity ρ itself to rise significantly as the line width shrinks —it is not a fixed value; rather, the finer the line, the larger ρ becomes.

Considering only the geometric shrinkage, the resistance has already increased by a factor of 13. After superimposing surface scattering and grain boundary scattering, the actual increase in resistance far exceeds this estimate based solely on the reduction in cross-sectional area [8][9].

---

6 Geometric Lock-In of Spatial Utilization and Signal Crosstalk Out of Control

Channel narrowing brings not only a drastic increase in resistance, but also a systemic collapse in spatial utilization and the loss of control over signal crosstalk. These two disasters share a single physical root—the cross-sectional area of interconnect channels cannot be proportionally compressed—yet this root cause has been systematically obscured by the mainstream narrative of "density scaling."

The routing resource consumed by an interconnect channel cannot be measured by line width alone; a complete two-dimensional cross-sectional perspective must be adopted. The physical space occupied by each channel in the cross-section is jointly determined by two directions: the sum of the line width w and the line spacing δ_w in the width direction, and the sum of the line height τ and the layer spacing δ_τ in the height direction. As the number of channels that must be accommodated per unit lateral surface area explodes by a factor of

approximately 13, w and τ are forced to shrink proportionally, but the spacings δ_w and δ_τ cannot keep pace—because their lower bounds are locked by insurmountable physical thresholds:

6.1 Dielectric Breakdown Field Strength

Sufficient insulating distance must be maintained between adjacent channels in both the width and height directions to prevent breakdown under the operating voltage. When the spacing is compressed to the nanometer scale, even a very low operating voltage can produce an electric field in the dielectric material strong enough to trigger leakage or breakdown. This is a hard constraint imposed by the intrinsic properties of the material and cannot be eliminated by process optimization.

6.2 Crosstalk Threshold and Signal Integrity Collapse

The coupling capacitance is proportional to the facing area of adjacent conductors and inversely proportional to the channel spacing. The facing area is jointly determined by the line height τ and the routing length—the larger τ is, the larger the facing area, the stronger the coupling capacitance, and the more severe the crosstalk [8]. This threshold defines a minimum permissible spacing in both the width and height directions. If

the spacing is forcibly compressed below the threshold, simultaneously switching signals will directly induce noise voltages exceeding the logic threshold, leading to an unacceptable data error rate. Suppressing crosstalk requires increasing the spacing or inserting shield lines, but this further compresses the already crowded routing space.

More fatally, the line height τ is locked from both sides here. Section 5 has already shown that to suppress the resistance surge, the industry is forced to maintain or even increase τ to preserve the cross-sectional area and reduce the per-unit-length resistance. However, increasing τ directly enlarges the facing area of adjacent conductors, raising the coupling capacitance and further worsening crosstalk. Increase τ and crosstalk intensifies; decrease τ and resistance soars—any adjustment of τ detonates one of the two disasters.

6.3 The Complete Two-Dimensional Cross-Sectional Lock-In

Combining the above constraints, the total routing resource consumption in the two-dimensional cross-section is $(w + \delta_w) \times (\tau + \delta_\tau)$. When the line width w is compressed far below the permissible $\delta_{w,min}$ in its direction, while the line height τ is pushed toward the process upper limit under resistance pressure but simultaneously locked by the crosstalk

threshold, the total routing resource becomes almost entirely dominated by the incompressible spacing components. The contribution of shrinking w and τ themselves to the overall density approaches zero. Spatial utilization therefore deteriorates catastrophically at the nanometer scale—directly contradicting the intuitive expectation that shrinking dimensions automatically improve density, yet this is an inescapable two-dimensional geometric penalty imposed by the fixed lateral surface area constraint.

This means that at the interconnect level, advanced nodes have already lost the possibility of mitigating problems by optimizing their own structure—the only way out is to transfer the contradictions from inside the chip to the more complex packaging interface. And the capacity of the packaging interface to bear channel narrowing is equally constrained by the physical limits demonstrated in Sections 8 through 11 of this paper.

---

7 Thermal Runaway and Signal Integrity Lock-In

When signals are transmitted through high-resistance channels, RC delay increases significantly, and signal amplitude suffers severe attenuation. A more fundamental problem lies in the fact that the drastic resistance increase creates a non-negligible heat source: Joule heat scales linearly with resistance, superimposing an additional thermal load on top of the already extremely high power density of advanced nodes. Rising

temperature causes carrier mobility to degrade, further increasing interconnect line resistance, forming the first positive feedback loop between heat and electricity.

The traditional compensation method is to insert repeaters along the channel, but this triggers a second positive feedback loop. Repeaters themselves consume dynamic power and generate Joule heat. The enormous number of channels necessitates a huge number of repeaters. The dense distribution of repeaters creates localized hotspots, where rising temperature increases leakage current, which in turn further increases repeater power consumption [8][10][11]. The superposition of these two positive feedback loops drives the temperature of the interconnect system irreversibly upward.

High temperature not only threatens the physical safety of the chip, but also directly dismantles signal integrity. Rising temperature alters the electrical characteristics of the dielectric material; the coupling capacitance drifts with temperature, and the amplitude and frequency characteristics of crosstalk noise change accordingly, transforming the noise from "predictable" to "uncontrollable." At the same time, the high temperature itself pushes the noise threshold to its limit, further compressing the design margin for crosstalk suppression measures—the

space for increasing line spacing or inserting shield lines has already been severely eroded by the heat.

Thermal runaway and signal integrity collapse are mutually causal and mutually locked, forming the most lethal positive-feedback closed loop in advanced-node interconnect systems: hotter→worse signal→more compensation→hotter. This is a physical deadlock that cannot be broken by external cooling or process optimization.

---

8 Mandatory Alignment Accuracy Imposed by Channel Narrowing

Advanced packaging and chiplet assembly do not eliminate the contradiction between the explosion in channel count and the limited interface area; they merely shift it to the packaging interface [2]. To accommodate a massive number of channels within that interface, interconnect pitch and line width must shrink further, directly forcing packaging processes to attain nanometer-scale alignment accuracy. Once the misalignment exceeds the tolerance of the channel width itself, the consequences will manifest in two irreversible forms.

A slight deviation may not cause complete misalignment, but it will

reduce the contact area, causing the contact resistance to soar—and resistance surge is precisely the starting point of the thermo-electric positive feedback chain demonstrated in Section 5 of this paper. A severe deviation is even more fatal: when interconnect channels are completely misaligned and signal lines connect to the wrong pins, at best, signals are interrupted and communication is lost; at worst, power and ground lines short together, and the entire chip is scrapped the instant it is powered on. This failure mode is not probabilistic, but an inevitable consequence of alignment accuracy being pushed to its physical limit.

At nanometer-scale spacing, any microscopic thermal expansion, vibration, or manufacturing tolerance is sufficient to trigger alignment failure. The multi-layer stacked structure of advanced packaging makes such failure, once it occurs, irreparable and irreversible—the entire package, along with all the expensive chips within it, is scrapped.

Therefore, the high-precision requirement of advanced packaging is not an optional technical choice, but a mandatory physical constraint imposed by channel narrowing. It pushes the process tolerance of packaging to the physically realizable limit, and directly converts any deviation beyond that limit into a system-level catastrophe.

---

9 2.5D Packaging: The Shifting and Transformation of Contradictions

2.5D packaging technologies, such as CoWoS, represent the mainstream industrial response to the above dilemmas. They shift the contradictions from within a single chip to the interposer and packaging interface. This shift is not without value—transforming the boundary of a problem is itself a core engineering strategy. However, it also introduces new physical challenges.

9.1 Complication of the Heat Dissipation Path

The introduction of chip stacking and silicon interposers forces heat flow to traverse multiple layers of material interfaces with relatively low thermal conductivity. Advanced nodes already have extremely high power density, and the packaging structure adds further thermal resistance. The nominal peak performance that users pay a high cost to obtain is often difficult to sustain under real-world continuous workloads due to hitting the thermal wall [2].

9.2 The Step-Change Increase in Packaging Cost

To stack interconnects within constrained space, the additional process

steps and ultra-high-precision equipment required push packaging cost from "negligible" to a magnitude comparable to wafer fabrication [6]. The marginal returns in thermal and signal integrity diminish even as costs escalate.

9.3 Thermomechanical Mismatch and Reliability Challenges

Silicon chips, silicon interposers, and organic substrates have different coefficients of thermal expansion. Under high-power operation, layers expand by different amounts, subjecting microbump connections to accelerated fatigue from cyclic thermal stress. Macro-level temperature control can regulate the average chip temperature, but localized hotspots generated by advanced-node power density produce heat fluxes approaching the thermal conductivity limits of silicon. It is micron-scale temperature gradients—rather than the average chip temperature—that are the root cause of thermomechanical mismatch failure [2]. This physical mechanism means that external cooling solutions alone cannot fundamentally eliminate the reliability risk.

Chip design is a systems discipline. When the single-point pursuit of extreme process nodes incurs disproportionate penalties in thermal, signal integrity, and mechanical reliability at the system level—penalties

ultimately borne by end users in the form of throttled performance and inflated costs—a reexamination of the technology roadmap becomes necessary.

---

10 3D Packaging: The Thermal Boundary of Vertical Stacking

Vertically stacked chips (3D packaging) attempt to bypass the edge interconnect bottleneck, but face even more severe thermal constraints. Heat generated in the middle layers of the stack must be conducted laterally to the edges for extraction, and the lateral thermal resistance increases linearly with distance. Under high power density, temperatures in the inner layers can easily exceed allowable limits [2].

The thermal predicament further triggers a structural reliability catastrophe. In a 3D package, the coefficients of thermal expansion (CTE) of the various chips, interposers, underfill, and substrate differ from one another. When the middle-layer chips heat up, their tendency to expand is strongly constrained by the upper and lower materials, generating significant internal stress. When heat accumulates in the middle layer due to impeded lateral conduction, the local temperature gradient further exacerbates this thermomechanical mismatch, causing the entire package to warp [2]. The stress induced by warpage is concentrated directly on the

microbumps—the most fragile connection points between the chips—leading to fatigue cracking of the solder joints, interfacial delamination, and, ultimately, the complete failure of the entire 3D package due to thermo-mechanical coupling effects. This is not a manufacturing defect, but an inescapable physical consequence of a multi-material stacking system under thermal load.

Therefore, for the foreseeable future, 3D stacking is more suitable for low-power memory or heterogeneous logic-plus-memory stacking, and is unsuitable for vertically stacking high-power logic chips within the same layer. Otherwise, it will face both an intractable thermal dead end and an intractable structural dead end simultaneously.

---

11 Hybrid-Node Chiplet: Physical Negative Synergy and Channel Mismatch

The hybrid-node chiplet approach currently being promoted by the industry retains the compute cores on an advanced node while migrating only the I/O and auxiliary functions to a mature node. This approach is packaged in the industry narrative as an ingenious architecture that "combines the best of both worlds." However, when examined through the lens of physical laws, it not only fails to eliminate the contradictions demonstrated in this paper, but instead causes the two process nodes to

mutually sabotage each other's inherent advantages, while exposing a fundamental dimensional mismatch in their interconnect channels.

The high power density and localized hotspots generated by advanced-node compute cores transmit thermal stress directly to the mature-node chips through the tightly coupled packaging structure, accelerating their aging and destroying the high reliability that mature nodes are supposed to provide [2]. Constrained by their own I/O bandwidth and latency, the mature nodes force the advanced-node compute cores to stall frequently and throttle their operating frequency, preventing their nominal performance from being realized in the system. The interconnect interface between the two demands extremely high alignment accuracy, yet the localized thermal expansion mismatch caused by the advanced nodes precisely concentrates stress at these microbump connections, subjecting the entire system to accelerated fatigue failure under thermal cycling [2].

A more fundamental problem lies in the inherent dimensional mismatch between the interconnect channels of the advanced and mature nodes. The interconnect line widths of the advanced node have been compressed to tens of nanometers, with pin pitches of only a few microns, requiring sub-micron alignment accuracy [8]. In contrast, the interconnect line

widths of the mature node are in the micrometer to sub-micrometer range, with pin pitches several to over ten times larger than those of the advanced node.

When these two types of chips must complete signal connections at the same packaging interface, only three outcomes are possible: the signal lines of the advanced node are forcibly widened, forfeiting their density advantage; the pins of the mature node are forcibly shrunk, increasing resistance and reliability risks; or a complex interposer adapter is introduced, further increasing cost and thermal resistance. This is not an engineering difficulty that can be optimized away—it is a fundamental physical scale mismatch between the two process nodes. Channel narrowing does not only occur inside advanced-node chips, but also at the interface where advanced and mature nodes must meet.

Therefore, the hybrid-node chiplet is not a combination of complementary strengths, but a case of physical negative synergy: the performance of the advanced node is diluted by the bandwidth limitations of the mature node, the robustness of the mature node is destroyed by the hotspots of the advanced node, and the dimensional mismatch between the two makes the interconnect interface the most vulnerable failure point in the entire system. This is a patchwork solution that shifts physical

contradictions from within a single die to the packaging interface at a higher cost, not a genuine paradigm shift [6].

---

12 Re-examining the Applicability Boundaries of Advanced Nodes

What the above analysis reveals is not the absolute conclusion that "advanced nodes cannot be used under any scenario." On the contrary, there exists a class of scenarios—where the pursuit of computational density overrides all else, and the system can afford the corresponding thermal and cost penalties—in which advanced nodes are the necessary choice.

What truly warrants caution is the industry's default reflex of treating advanced nodes as the standard option for every scenario. When scenarios with stringent power constraints, relaxed area requirements, or extremely high reliability demands are also blindly pushed onto the process-scaling race track, the system is forced to cover the warnings of physical laws with layer upon layer of engineering patches. Eventually, these warnings are passed on to end users in the form of soaring costs and discounted performance [6].

More worthy of reflection is the fact that as process nodes approach ~1

nm, the triple loss of control stemming from quantum tunneling [3], statistical fluctuation [3], and interconnect narrowing [8] superimposes, causing process "uncontrollability" to overwhelm "usability." Chips can still be manufactured, but their distributions of yield, frequency, and power dissipation become statistically unpredictable [3]. To extract the last bit of nominal performance, manufacturers often pay a disproportionate price in power consumption [11]. In scenarios of sustained operation under a fixed power budget, a more power-efficient and physically robust "next-best" node may actually deliver higher practically usable total compute. This means advanced nodes are not only wasteful in misapplied scenarios; even within their supposed target domains, a divergence between nominal performance and practically deliverable performance has begun to set in. When uncontrollability eclipses usability, a so-called "more advanced" node has fallen into a utility inversion for the majority of practical engineering contexts.

---

13 Mature-Node Chiplets: The Physical Return of System Integration

The above analysis also reveals a fact that has been systematically obscured by the industry narrative: channel narrowing is a predicament unique to advanced nodes. Mature nodes have low transistor density, so the number of I/O channels required per unit area is naturally smaller, and

channel widths need not be forced down to the nanometer scale [8].

Here it is necessary to highlight a fundamental physical advantage that mature nodes possess with respect to interconnect delay. The RC delay of an interconnect line is proportional to the product of its per-unit-length resistance and capacitance. At advanced nodes, after interconnect lines are compressed to nanometer-scale cross-sections, the extremely small cross-sectional area causes the per-unit-length resistance to soar; dense repeater insertion is then mandatory to barely maintain delay, which in turn directly drives up power consumption and thermal density [8][11]. In contrast, mature-node interconnect lines typically have widths in the micrometer to sub-micrometer range, with cross-sectional areas dozens of times larger than those of advanced-node thin lines. Their per-unit-length resistance is thus more than an order of magnitude lower [8][9]. Such wide lines can maintain an acceptable RC delay over longer distances without any repeaters, and they do not introduce the extra power consumption and localized hotspots that repeaters bring [10]. This is precisely the irreplaceable physical capital of mature nodes in the context of system integration.

At the same time, mature nodes have low power density, no local hotspot hazards, and greatly reduced thermomechanical mismatch risks [2]. They

require no sub-micron alignment and can be served by traditional packaging. Small-die assembly circumvents the single-die yield trap [5][7], and the low resistance of individual interconnect lines keeps signal-transmission power consumption manageable and economically viable [6].

This is not an argument for comprehensively replacing advanced nodes with mature ones. Rather, it points to a physically more self-consistent system integration path. The mature-node chiplet approach argued for in this paper requires that the compute cores themselves must also adopt Mature Nodees, so as to eliminate interconnect narrowing—the predicament unique to advanced nodes—at its physical root. Do not pursue extreme single-point density; pursue system-level physical self-consistency: every interconnect channel is sufficiently wide, RC delay is naturally manageable, every heat source is amply dispersed, and every packaging interface is spared the cyclic tearing of thermal expansion.

This is not a conservative retreat, but a renewed respect for physical laws. At a time when the physical penalties of advanced nodes are escalating rapidly and their practically deliverable performance is approaching a utility boundary, returning to system-level rational trade-offs is not only a

response to engineering ethics, but also a responsibility to user interests and energy efficiency [6]. When one road becomes narrower and narrower, stepping back and choosing a broader path is not defeat; it is wisdom.

References

[1] G. E. Moore, "Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff.," in *IEEE Solid-State Circuits Society Newsletter*, vol. 11, no. 3, pp. 33-35, Sept. 2006, doi: 10.1109/N-SSC.2006.4785860.

[2] H. Kim, J. Y. Hwang, S. E. Kim, Y. -C. Joo and H. Jang, "Thermomechanical Challenges of 2.5-D Packaging: A Review of Warpage and Interconnect Reliability," in *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 13, no. 10, pp. 1624-1641, Oct. 2023, doi: 10.1109/TCPMT.2023.3317383.

[3] M. Salmani Jelodar et al., "Tunneling: The major issue in ultra-scaled MOSFETs," 2015 IEEE 15th International Conference on Nanotechnology (IEEE-NANO), Rome, Italy, 2015, pp. 670-673, doi: 10.1109/NANO.2015.7388694.

[4] R. H. Dennard, F. H. Gaensslen, H. -N. Yu, V. L. Rideout, E. Bassous and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," in *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, pp. 256-268, Oct. 1974, doi: 10.1109/JSSC.1974.1050511.

[5] C. H. Stapper, "On Murphy's yield integral (IC manufacture)," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 4, no. 4, pp. 294-297, Nov. 1991, doi: 10.1109/66.97812.

[6] A. Mallik et al., "Economics of semiconductor scaling - a cost analysis for advanced technology node," 2019 Symposium on VLSI Technology, Kyoto, Japan, 2019, pp. T202-T203, doi: 10.23919/VLSIT.2019.8776521.

[7] S. Naffziger et al., "Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families : Industrial Product," 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 2021, pp. 57-70, doi: 10.1109/ISCA52012.2021.00014.

[8] R. Brain, "Interconnect scaling: Challenges and opportunities," 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco,

CA, USA, 2016, pp. 9.3.1-9.3.4, doi: 10.1109/IEDM.2016.7838381.

[9] W. Wu, S. H. Brongersma, M. Van Hove and K. Maex, "Influence of surface and grain-boundary scattering on the resistivity of copper in reduced dimensions," in *Applied Physics Letters*, vol. 84, no. 15, pp. 2838-2840, Apr. 2004, doi: 10.1063/1.1703844.

[10] J. C. Ku and Y. Ismail, "Thermal-Aware Methodology for Repeater Insertion in Low-Power VLSI Circuits," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 15, no. 8, pp. 963-970, Aug. 2007, doi: 10.1109/TVLSI.2007.900749.

[11] P. Kapur, G. Chandra and K. C. Saraswat, "Power estimation in global interconnects and its reduction using a novel repeater optimization methodology," Proceedings 2002 Design Automation Conference (IEEE Cat. No.02CH37324), New Orleans, LA, USA, 2002, pp. 461-466, doi: 10.1109/DAC.2002.1012669.