Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Syste (CONF-SEML 2026)
DOI: 10.54254/2755-2721/2026.GU35086

A Review of Quantitative Decomposition Studies on the Drivers of Microprocessor Performance Growth

Changyuan Ma

Leeds College, Southwest Jiaotong University, Chengdu, China
1664094296@qq.com

Abstract. In the post-Moore's Law era, microprocessor performance growth continues to slow, as the traditional approach of process scaling and frequency increases has reached a bottleneck. Physical limits, power constraints, and architectural innovation have become the three core factors influencing performance growth. However, academia disagrees on quantifying their individual contributions and lacks a unified framework for analyzing their coupled effects. This paper systematically reviews classical theories and research paradigms related to microprocessor performance growth. The study identifies three core consensuses: power constraints are the primary limiting factor throughout the performance growth cycle; architectural innovation is the key driver in the post-Moore era; and the contribution of process scaling shows clear diminishing returns. The study further compares the mechanisms of physical limits and power constraints with quantitative research on architectural innovation. By examining mainstream quantitative decomposition methods for these three factors, it highlights core shortcomings in existing research and outlines future directions. This paper provides a theoretical foundation for research on microprocessor architectural innovation and power optimization in the post-Moore's Law era.

Keywords: microprocessors, performance scaling, physical limits, power constraints, architectural innovation, quantitative decomposition

1. Introduction

Since Moore's Law was proposed in 1965, microprocessor performance has grown exponentially over decades, serving as the core engine driving the evolution of the information technology (IT) industry. However, beginning in the mid-2000s, Dennard's scaling law gradually ceased to hold, and single-core frequencies hit the "frequency wall" at approximately 4.0 GHz [1]. Process technology evolution has also encountered physical limits such as quantum tunneling and surging leakage current, while soaring manufacturing costs for advanced processes have propelled microprocessor development into the post-Moore's Law era. Architectural innovation and power optimization have gradually replaced process scaling as the primary pathways for performance growth, and power constraints have shifted from secondary considerations to core prerequisites for architectural design. Against this backdrop, the academic community has conducted extensive research on the three key factors: physical limits, power constraints, and architectural innovation.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU35086

However, there are discrepancies in the results of performance contribution measurements across different studies, and there remains a gap in the quantitative analysis of the coupled effects of multiple factors; a unified quantitative decomposition framework has yet to be established.

Accordingly, this study focuses on general-purpose desktop and server-class microprocessors from leading manufacturers between 1985 and 2025. By systematically reviewing the relevant research on these three core factors, this study fills a gap in the current literature. This study provides an academic reference for selecting R&D paths for microprocessor architectures and formulating industrial technology strategies in the post-Moore's Law era.

2. Theoretical foundations and research paradigms of microprocessor performance growth drivers

2.1. Classical theories of performance growth

The core theoretical foundations of microprocessor performance growth stem from Moore's Law and Dennard's scaling law. The academic community widely acknowledges that these two laws have jointly underpinned the rapid development of microprocessors over several decades, forming the core framework for early performance growth research [1]. Microprocessor performance growth is fundamentally rooted in Moore's Law and Dennard's scaling law. Moore's Law observed that transistor density doubles every 18–24 months [2]. Dennard's scaling law stated that transistor size, voltage, and current scale proportionally, keeping power density stable [1]. Together, they enabled simultaneous increases in transistor density and frequency. Between 1985 and 2004, single-core frequency grew at 28.3% annually [3]. However, after the mid-2000s, Dennard scaling broke down and Moore's Law's marginal benefits diminished, making the traditional process-and-frequency scaling path unsustainable.

2.2. Evolution of the multi-factor-driven research paradigm

Along with the shift in microprocessor technology pathways, academic research on the drivers of performance growth has also adjusted. Early research paradigms centered on a single-process-driven approach, treating process scaling as the sole factor for performance growth. Research focused primarily on process miniaturization, viewing architectural design and power control merely as supplementary measures without addressing their independent roles. Following the emergence of the "frequency wall" in 2004, the paradigm shifted toward multi-factor synergy: achieving performance gains through architectural innovation under power constraints, with the research scope expanding to include the coupled system of process, architecture, and power.

Furthermore, the academic community gradually refined core theoretical models: Amdahl's model quantified the performance ceiling of parallel architectures, clarifying the constraint of serial code on multi-core performance and becoming the theoretical foundation for multi-core research. Pollack noted that performance scales with the square root of architectural complexity, providing a reference for calculating marginal benefits of architectural innovation. Tebaldi and Zanari proposed a unified analytical approach for power efficiency of physical systems [4].

DOI: 10.54254/2755-2721/2026.GU35086

3. Literature review on physical limits and power consumption constraints

3.1. Constraints on performance growth imposed by physical limits

3.1.1. Short-channel effects and quantum tunneling

Physical limits are the core constraint on process scaling and the fundamental limitation on performance growth in the post-Moore's Law era. Current research generally agrees that the physical limits of silicon-based transistors primarily stem from the short-channel effect and quantum tunneling. When the channel length shrinks to a certain threshold, the width of the source-drain barrier becomes too narrow, allowing electrons to directly pass through the barrier via quantum tunneling. This prevents the transistor from turning off properly, resulting in an exponential increase in leakage current. Simultaneously, the short-channel effect triggers issues such as threshold voltage drift, which further reduces transistor reliability. Bohr and Young's research reveals that quantum tunneling and leakage current act as the main bottlenecks for process nodes below 22 nm [5]. Traditional planar transistor structures are no longer capable of meeting the demands of advanced processes.

3.1.2. Threshold estimation and physical limits

To estimate thresholds associated with physical limits, Fan Heng et al. conducted total dose effect tests on commercial microprocessors from the same manufacturer with feature sizes of 180 nm, 90 nm, and 40 nm. The results showed that the irradiation failure thresholds for the three processor sizes were 331±36.28 Gy(Si), 355.5±41.51 Gy(Si), and 365.28±20.15 Gy(Si), respectively. The reduction in feature size led to a continuous increase in the irradiation sensitivity of the device core, and the reliability threshold of transistors gradually decreases with process scaling, indirectly confirming the intensifying physical constraints resulting from size reduction [6]. It is widely accepted in the academic community that the physical limit of the channel length for traditional silicon-based planar MOSFETs is around 20 nm. FinFET structures can push this limit to 3 nm, while GAA structures can further extend it to 2 nm. Beyond this node, transistors will lose their normal switching capability.

3.1.3. New materials and structures

To address fundamental physical limits, current research focuses on developing new materials and device structures. Li and Lanza optimized the process through 130 tests and made wafer-scale 2D material transistors. Their channel length can be less than 1 nm. This provides a new mass-producible way to break traditional silicon transistors' physical size limits [7]. Concurrently, novel packaging structures such as 3D stacking and chip-on-chip (CoC) have emerged as key approaches to circumventing physical limits. Techniques like 3D stacking do not require reducing the size of individual transistors but instead increase transistor density through vertical integration. This serves as a vital technological pathway for maintaining performance advancement in the post-Moore's Law era.

DOI: 10.54254/2755-2721/2026.GU35086

3.2. Mechanisms and quantitative assessment of power constraints

3.2.1. Power constraints

Power constraints are a core premise of processor design in the post-Moore's Law era. The academic community generally classifies microprocessor power consumption into dynamic and static power, with significant differences in their constraint mechanisms. Dynamic power stems from the charging and discharging of capacitance during transistor switching; its magnitude is proportional to the square of the operating frequency and voltage. This implies that frequency increases lead to exponential growth in power consumption. The core reason for the emergence of the "frequency wall" in 2004 was that dynamic power exceeded the upper limit of the chip's thermal dissipation capacity. Static power consumption stems from leakage current when transistors are turned off. Its magnitude rises rapidly as process nodes shrink. For FinFET technologies below 22 nm, static power consumption accounts for more than 42% of the total processor power consumption under full load, becoming the dominant contributor to power rise in advanced manufacturing processes [6].

3.2.2. Quantitative comparison of "frequency wall" thresholds

Regarding the frequency wall thresholds directly governed by power constraints, researchers have obtained multiple sets of quantitative verification results. ll data are derived from the original measurement results in the corresponding references. Based on the literature, mainstream general-purpose processors have a single-core frequency threshold of 3.8-4.0 GHz, corresponding to a power density threshold of 10-11 W/mm². Mainstream manufacturers reached this range by 2004, leading to a shift from single-core frequency scaling to multi-core parallel processing [3, 8]. In contrast, heterogeneous computing processors can relax frequency constraints on general-purpose cores by deploying dedicated accelerator cores, lifting the power density limit to over 14 W/mm² and thereby extending performance improvement potential [9].

3.2.3. Optimization studies under power constraints

Scholars have quantitatively studied the effects of power constraints. Antonio del Vecchio et al. quantified how communication interface latency affects the efficiency of end-to-end power management schemes for high-performance processors, showing that communication latency directly reduces power control response speed and thus limits performance [9]. Tengyue Pan et al. evaluated the impact of thermal efficiency on processor temperature via finite element simulation, finding that rising ambient temperatures significantly increase core temperatures, while optimizing component layout can lower the processor's operating temperature by about 10 °C, confirming the direct impact of thermal performance on power constraints [8]. Lian proposed a design method based on dynamic timing margin compression. It improves performance by 31% without increasing power, or reduces power by 36% with only 0.6% area overhead. This verifies that architectural optimization can reduce power constraints [10].

DOI: 10.54254/2755-2721/2026.GU35086

4. The performance-driving effects of architectural innovation

4.1. Quantifying the effects of classical architectural innovation

4.1.1. Instruction-level parallelism (ILP)

Against the backdrop of diminishing returns from process scaling, architectural innovation has become the primary driver of performance growth. Research on conventional processor architectures falls mainly into three categories: instruction-level parallelism, thread-level parallelism, and memory hierarchy optimization, each corresponding to a separate technological evolution stage.

Core technologies for instruction-level parallelism (ILP) optimization include superscalar processing, out-of-order execution, and branch prediction. These were the primary sources of performance growth, excluding process technology, from 1990 to 2004. Quantitative analysis by Hennessy and Patterson shows that ILP optimization contributed to an average annual growth rate of 52% in single-threaded performance from 1990 to 2000 [3]. However, ILP optimization faces a clear efficiency ceiling. Simulation results show that for general-purpose processors, when the instruction issue width exceeds 8 instructions per cycle, the performance degradation arising from higher branch prediction error rates fully offsets the IPC benefits brought by a wider issue width, and single-threaded performance stagnates as the issue width further increases [3].

4.1.2. Thread-level parallelism (TLP)

The core of thread-level parallelism (TLP) optimization is the multi-core architecture. It was the primary source of performance growth from 2004 to 2014. From 2004 to 2020, mainstream server processor core counts rose from 2 to 64, with a 35.7% annual compound growth rate. Verification via a panel data fixed-effects model demonstrates that the marginal coefficient of core count on multithreaded performance rose from 0.23 to 0.67, thus establishing multicore architectures as the primary driver of performance growth [3]. The academic community generally adopts Amdahl's Law to quantify the performance ceiling of multi-core architectures.

4.1.3. Storage system optimization

Core technologies for memory hierarchy optimization include cache hierarchy expansion and interconnect architecture optimization. These are primarily used to address the "memory wall" problem and reduce memory access latency. Quantitative studies on the memory hierarchy of general-purpose processors show that, under the same process technology and architecture, doubling the capacity of the last-level cache can reduce average processor memory access latency by 32%-41%, corresponding to an 11%-16% increase in single-threaded integer performance [11]. Research by Wu Tiebin et al. on core computing architectures for exascale computing also confirms the importance of co-optimizing the memory hierarchy and computing architecture [11].

4.2. Potential and challenges of emerging architectures

In recent years, the academic community has conducted extensive cutting-edge research on emerging architectures. The core objective is to circumvent the physical and power constraints of traditional architectures. Current key research directions include heterogeneous computing, compute-in-memory (CIM), chiplet-based architectures, and open-source instruction set architectures (ISAs).

DOI: 10.54254/2755-2721/2026.GU35086

4.2.1. Heterogeneous computing

Heterogeneous computing is currently the most commercially mature direction. Its core lies in combining general-purpose cores with specialized accelerator cores. Alejandra Sanchez Flores and colleagues designed a neural network accelerator based on pulsed arrays. Their research demonstrated that heterogeneous architectures can break through the energy efficiency bottlenecks of general-purpose processors in AI inference scenarios [12]. Yanhan Huang's research also proposed an acceleration scheme combining ARM and FPGA [13].

4.2.2. Memcomp

Compute-in-memory architectures are a cutting-edge solution to the "memory wall" problem. Traditional von Neumann architectures have high power consumption and latency from data movement. In contrast, compute-in-memory architectures integrate computing and storage units, enabling computations to be performed directly within the memory array. Multiple studies have shown that SRAM-based compute-in-memory architectures for AI inference scenarios can achieve up to a 12-fold improvement in energy efficiency compared to traditional architectures [12]. However, CIM architectures have major limitations: Memory cell characteristics limit computational accuracy. The architecture also lacks generality and only works for specific computational patterns.

4.2.3. Chiplet architecture

Chiplet-based architecture is a key direction for overcoming the fundamental physical limits and cost constraints of single-chip designs. Gu Kai et al. point out that chiplet technology enables the integration of bare chips manufactured using different processes through packaging [14]. It overcomes the size limitations of a single chip and reduces design and manufacturing costs. The core challenge lies in the fact that interconnections between chiplets introduce additional latency and power consumption overhead. Overall, these emerging architectures avoid simple process scaling and are a key post-Moore era performance growth direction. However, they still face challenges like limited versatility and high interconnection overhead.

5. Quantitative decomposition methods for driving factors and future prospects

5.1. Typical quantitative decomposition methods

To quantify the three factors' contributions, scholars have developed three main quantitative models: regression analysis, factor decomposition, and machine learning-assisted methods. Regression analysis controls for confounders (e.g., time trends, product positioning) to estimate marginal net impacts. It suits long-term, large-sample studies but cannot quantify cumulative contribution rates. Factor decomposition models offer residue-free decomposition, handle negative values, and measure dynamic contributions by phase. However, they require high data completeness and are unsuitable for small samples. Machine learning-assisted models handle nonlinear coupling and dynamic loads but lack interpretability and need large experimental datasets [15]. Based on these methods, key quantitative results on full-cycle performance contributions are summarized in Table 2.

DOI: 10.54254/2755-2721/2026.GU35086

Table 1. Quantitative comparison of contribution rates of factors driving microprocessor performance growth [3, 5]

Sample period	Core Quantitative Methods	Cumulative contribution of process scaling /%	Cumulative contribution rate of architectural innovation /%	Cumulative contribution to power consumption constraints (negative) /%
1985—2015	Performance Breakdown Model	38.1	32.6	29.3
1995—2015	Calculation of Marginal Returns to Scale	41.2	28.7	30.1

Note: All data are derived from full-cycle projections in the corresponding references.

Over the period from 1985 to 2020, the contributions of process scaling, architectural innovation, and power constraints to general-purpose processor performance growth were broadly balanced. The main disagreement lies in the post-2010 period, where estimates of architectural innovation's contribution vary by up to 12.5 percentage points across studies. This discrepancy stems from whether ARM's low-power architecture is adopted and whether heterogeneous computing is included in architectural innovation. Overall, all methods have limitations, and a unified quantitative decomposition framework is yet to be established.

5.2. Future directions

Future work should develop a multi-factor, coupled, quantitative decomposition framework that captures both independent and synergistic contributions. Meanwhile, establishing dynamic scenario models covering AI and cloud computing via machine learning and hardware simulation. Refine evaluation systems for emerging architectures using unified benchmarks and real-chip data, and incorporate industrial and economic factors such as manufacturing costs and R&D investments.

6. Conclusion

This paper finds that microprocessor performance growth in the post-Moore era has shifted from relying solely on process scaling and frequency increases to a multi-factor synergy model. Existing research agrees that power constraints are the core limiting factor, architectural innovation is the key driver, and process scaling shows diminishing returns. In 2004, single-core frequencies reached the 3.8–4.0 GHz "frequency wall" due to the combined effects of physical limits, power constraints, and architectural design. It identifies three challenge: quantification of multi-factor coupling effects, adaptability to dynamic scenarios, and evaluation frameworks for emerging architectures. Future research should establish a multi-factor coupled quantitative decomposition framework, refine evaluation systems for dynamic scenarios and emerging architectures, and expand research perspectives to multiple dimensions, thereby providing theoretical and empirical support for microprocessor technology evolution in the post-Moore's Law era.

However, due to limitations in existing literature and reliance on secondary data, this study did not conduct actual microprocessor testing. Therefore, the generalizability of its conclusions requires further validation. Future efforts should focus on establishing standardized, publicly accessible performance databases and designing reproducible multi-factor decomposition models.

DOI: 10.54254/2755-2721/2026.GU35086

References

[1] Dennard, R. H., et al. (1974). Design of ion-implanted MOSFETs with very small physical dimensions. IEEE Journal of Solid-State Circuits, 9(5), 256-268.
[2] Moore, G. E. (1965). Cramming more components onto integrated circuits. Electronics, 38(8), 114-117.
[3] Hennessy, J. L., & Patterson, D. A. (2017). Computer architecture: A quantitative approach (6th ed.). Morgan Kaufmann.
[4] Tebaldi, D., & Zanasi, R. (2024). A unified methodology for the power efficiency analysis of physical systems. Journal of the Franklin Institute, 361(1), 1=28.
[5] Bohr, M. T., & Young, I. A. (2017). CMOS scaling trends and beyond. IEEE Micro, 37(6), 20-29.
[6] Fan, H., Liang, R., Chen, F., et al. (2025). Experimental study on the total dose effects of microprocessors with different feature sizes. Microelectronics, 55(1), 59-64.
[7] Li, X., & Lanza, M. (2026). Building two-dimensional microprocessors with a foundry-inspired strategy. Nature Electronics, 9(2), 112-121.
[8] Pan, T., et al. (2025). Thermal performance analysis and structural optimization of main functional components of computers. Applied Sciences, 15(17).
[9] del Vecchio, A., et al. (2024). Performance characterization of hardware/software communication interfaces in end-to-end power management solutions of high-performance computing processors. Energies, 17(22).
[10] Lian, Z. (2024). Research on high-performance processor design techniques based on dynamic timing margin compression [Doctoral dissertation, Shanghai Jiao Tong University].
[11] Wu, T., Guo, F., & Wang, D. (2023). Research progress on high-performance processor core computing architectures for exascale computing. Computer Engineering and Science, 45(5), 761-771.
[12] Sanchez Flores, A., et al. (2024). Energy and precision evaluation of a systolic array accelerator using a quantization approach for edge computing. Electronics, 13(14).
[13] Huang, Y. (2026). Design of a microprocessor equipped with a convolutional neural network acceleration module. Modern Information Technology, 10(1), 17-21.
[14] Gu, K., Fang, J., Liu, T., et al. (2025). Research progress on processor-integrated chip architectures. Intelligent Security, 4(4), 102-112.
[15] Zhai, J., Ling, Z., Bai, C., et al. (2024). A review of machine learning-assisted microarchitecture power modeling and design space exploration. Computer Research and Development, 61(6), 1351-1369.