From Material Properties to System Architecture: Exploring Collaborative Pathways for Semiconductor Innovation in the Post-Moore Era

Haotian Liu*

Ulink High School of Suzhou Industrial Park, Jiangsu, 215000, China

*Corresponding author: c37673031@gmail.com

Abstract. When the traditional silicon-based semiconductor chip industry encounters bottlenecks caused by physical limits and can no longer follow Moore's Law—that is, when it enters the post-Moore era—how should it move forward? This paper discusses why the chip industry cannot continue its linear process scaling, taking quantum tunneling (e.g., electron tunneling), Compton scattering, and economic factors as examples. It then explores corresponding solutions from the perspectives of packaging, architecture, and materials. In terms of packaging, the potential of Fan-Out Panel-Level Packaging (FOPLP) with larger panels, as well as its current challenges, is examined. From an architectural standpoint, optical computing— which leverages photonics—is analyzed, along with its potential future applications based on its inherent characteristics. Additionally, neuromorphic computing—inspired by the operation of human neurons—is discussed, highlighting neuromorphic computing's advantages and limitations. Regarding materials, the speed advantages of two-dimensional materials (e.g., graphene) are analyzed, along with the potential innovations brought by topological insulators and spintronics. This paper thoroughly examines the challenges currently facing the semiconductor industry from multiple perspectives and proposes several potential solutions, thereby providing significant guidance for breaking through the industry's current impasse.

Keywords: Moore's Law; Chip Industry; Manufacturing Process.

1. Introduction

Since Moore proposed at the 1975 IEEE International Electron Devices Meeting (IEDM) that the number of transistors integrated on semiconductor chips doubles every two years, the entire chip industry has undergone a complete transformation. At least until 2000, Moore's Law largely guided the entire chip industry. However, due to various reasons, this pivotal law will come to an end in the coming years, and the industry will inevitably enter the post-Moore era [1].

While there are multiple reasons for the limitations in chip process scaling, the quantum tunneling effect has long been widely regarded as one of the most critical issues. According to the tunneling probability formula $T \approx e^{-2\kappa L}$, as the thickness L of the transistor's gate dielectric decreases, the tunneling probability T increases exponentially. This means that as transistors shrink, the tunneling leakage current rises dramatically. Similarly, based on Heisenberg's Uncertainty Principle ($\Delta x \cdot \Delta p \geq 2\hbar$), when the process reaches the quantum scale, requiring extremely low uncertainty in electron position (Δx), this leads to significant uncertainty in electron momentum (Δp), making quantum tunneling inevitable [2].

More importantly, the pursuit of further process scaling is inherently constrained by physical limits. Irreversibly writing a single bit requires a minimum energy of $E_{bit} = k_B T \ln 2$ (where k_B is Boltzmann's constant, and T is temperature). At room temperature (300 K), this corresponds to approximately 0.017 eV per operation. Below this threshold, noise and thermodynamic entropy will render reliable switching impossible [3]. To accurately measure the position of a particle, photons with shorter wavelengths (and thus higher energy) are required. However, if the photon's wavelength is shorter than the electron's Compton wavelength, the interaction between the photon and electron is no longer a simple position measurement. Instead, it results in Compton scattering ((($\Delta\lambda = \lambda -$

$\lambda 0 = \frac{h}{m_e c} - \cos\theta$), where h is Planck's constant, m_e is the electron rest mass, c is the speed of light, and θ is the scattering angle) and the generation of new particles [4].

In addition, there are secondary issues, such as the technical challenges faced by major enterprises when replacing transistor architectures and the impact of interface oxides at the nanoscale. From an economic perspective, the narrowing of marginal economic benefits and investment return windows has led to fewer enterprises being willing to make large-scale investments. These issues will be discussed in greater detail later in this paper.

This paper will explore potential solutions from three perspectives: packaging technology, architectural design, and materials. In terms of packaging technology, Fan-Out Wafer-Level Packaging (FOWLP) has already become an indispensable part of the chip industry; however, people still know very little about Fan-Out Panel-Level Packaging (FOPLP). The two technologies are fundamentally similar in principle but differ in their base substrates: FOWLP uses wafers, while FOPLP uses larger-sized panels. This panel-based design makes FOPLP more cost-effective and gives it greater potential for enhancing computing power. Its thinner profile also makes it more suitable for 3D stacking technology [5].

There are also numerous architectural innovations that fundamentally disrupt existing computing paradigms, rather than simply replacing a few electronic components. Optical computers are considered one of the strongest contenders. Photons inherently travel much faster than electrons and have almost no resistive loss, resulting in far less heat dissipation and, in theory, significantly lower energy consumption [6]. Another more radical architecture is neuromorphic computing. By mimicking the human brain, it overcomes the von Neumann bottleneck through an in-memory computing framework. It integrates processing units (simulated neurons) and storage units (simulated synapses)—just like the human brain—and only emits spikes when the received pulse signal reaches a certain threshold. This is known as a Spiking Neural Network (SNN) [7].

From a materials standpoint, two-dimensional (2D) materials such as graphene are gradually gaining attention. 2D materials consist of a single layer of atoms, which greatly reduces electron scattering in three-dimensional space, thereby effectively increasing electron conduction speed [8]. Even more revolutionary is the use of topological insulators to advance spintronics, which can significantly boost transmission speeds. Topological insulators are insulating in their interior, so electrons can only propagate along their surfaces. Here, electron spin and momentum are locked together, eliminating electron scattering or collisions entirely. Unlike traditional electronics, spintronics uses the spin of electrons rather than their charge to represent information—and topological insulators are inherently excellent spin-polarized channels. In this ideal model, information transmission would combine low power consumption, high speed, and non-volatility [9]. This paper aims to discuss the various challenges currently faced by traditional silicon-based CMOS technology, including physical limits, chip manufacturing hurdles, and economic factors. It also attempts to propose innovative solutions from the perspectives of packaging technology, system architecture, and materials science.

2. Problems Encountered in Precision Chip Processing

2.1. Quantum Tunneling

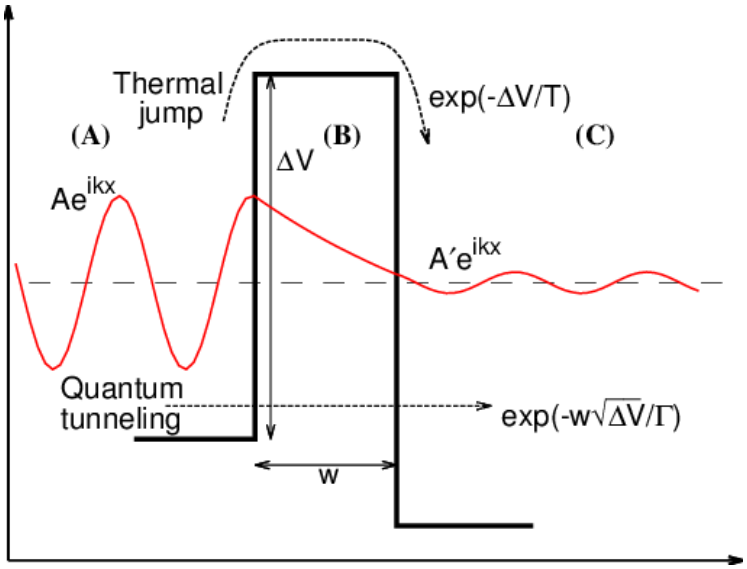


Fig.1 Quantum tunneling across a potential barrier by thermal hopping [10]

In traditional classical physics, a particle is regarded as having mass and therefore requires sufficient energy to pass through the potential barrier of an insulator—a process known as Thermal Jump. But in quantum mechanics, matter exhibits wave-particle duality. Its wave function does not drop to zero at the edge of the potential barrier; instead, it decays exponentially inside the barrier. Thus, on the other side of the potential barrier, the wave function still has a non-zero amplitude (as shown in Fig. 1), which means the particle still has a certain probability of tunneling through. This process is called quantum tunneling [11].

In the semiconductor chip industry, quantum tunneling primarily occurs in the gate oxide. As chip process scaling advances, the thickness of the oxide layer has been thinned to the quantum scale, making quantum tunneling more likely. Moreover, during this process, even a small change in voltage will cause a very large change in current, leading to a significant increase in gate leakage current. This negatively impacts the chip's performance, power consumption, heat generation, and reliability.

2.2. Compton Wavelength

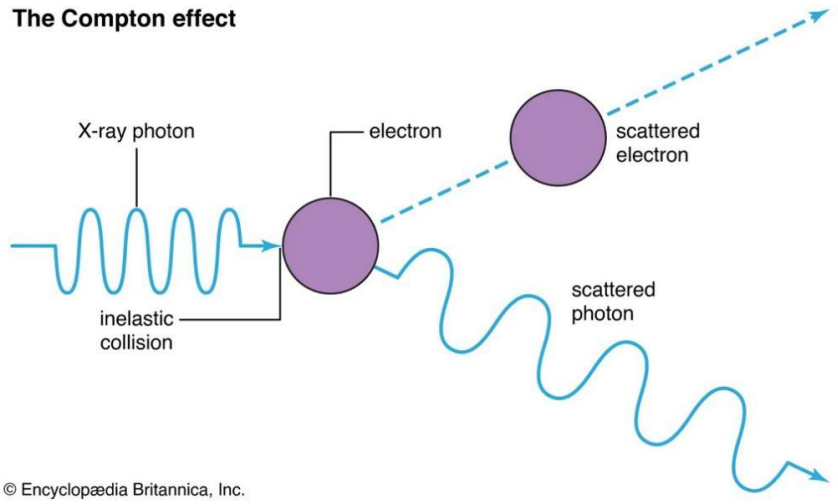


Fig.2 Compton scattering [12]

More broadly, this can be discussed directly from the perspective of the fundamental length associated with the intrinsic quantum properties of stationary particles—namely, the Compton

wavelength. From this viewpoint, the phenomenon most relevant to the chip industry is Compton scattering. Compton scattering describes the process by which high-energy photons (e.g., X-rays or gamma rays) lose energy when interacting with matter, as shown in Fig. 2.

In the context of the semiconductor chip industry, when chip process scaling becomes sufficiently advanced (especially when the thickness of the gate oxide layer decreases) to approach the Compton wavelength of the electron ($2.4263102367(11) \times 10^{-12}$ m), the quantum behavior of electrons (e.g., wave-particle duality) becomes particularly pronounced.

From a quantum perspective, the Compton wavelength provides a fundamental scale that defines the characteristic dimensions of the electron when treated as a wave. As the process scale approaches this fundamental physical length, classical physics models break down, and quantum tunneling becomes inevitable [4, 13].

2.3. Economic Factors

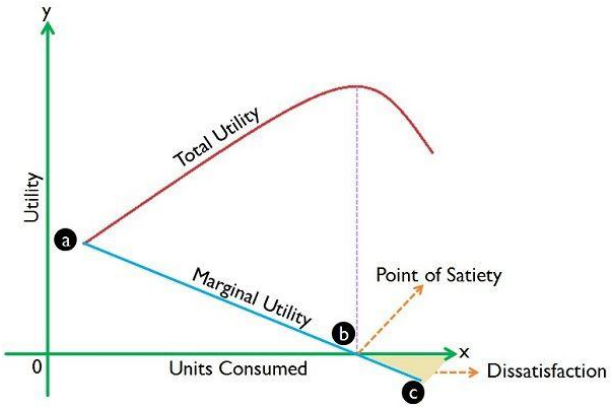


Fig.3 Total utility curve [14]

Some industry insiders seem overly optimistic about the positive linear relationship between chip process scaling and economic returns, and they underestimate the impact of marginal economic utility [15].

Fig. 3 illustrates Total Utility and Marginal Utility clearly and unambiguously. Let's analyze Fig. 3 in detail: the Total Utility curve rises first, indicating that as "consumption" (here referring to investment in chip process advancement) increases, Total Utility also increases. It then reaches a saturation point, where Total Utility is maximized, after which it begins to decline. Turning to the Marginal Utility curve, Marginal Utility decreases continuously—a reflection of the Law of Diminishing Marginal Utility. It drops to zero at point b and becomes negative thereafter. This phenomenon aligns with the trend that Total Utility peaks and then declines, and consumer demand for further process advancement starts to weaken.

This trend is particularly evident in the semiconductor industry: processes are growing more complex, chip manufacturing is becoming harder to fine-tune, design cycles are lengthening, and investment windows are being forced to expand. However, consumer demand for more advanced (complex) chips is diminishing, and the return on investment is disproportionate to the massive inputs [16].

2.4. Other Issues

There are also some relatively less prominent reasons (for chip process limitations) that are still worth mentioning. Semiconductor manufacturing is a global industry. As chip fabrication processes continue to advance, this progress poses not only challenges in addressing individual technical issues but also tests the collaborative capabilities of the global supply chain. As chip manufacturing processes become more refined, even seemingly minor variations in the thickness of the gate oxide layer can have a significant impact on chip performance and reliability.

3. FOPLP

In recent years, there have been growing calls to boost production and reduce costs. The transition from traditional FOWLP (using 300 mm wafers) to more advanced packaging technologies—specifically FOPLP—is almost inevitable. FOWLP has already delivered significant value in the global chip industry, but its costs remain relatively high, leaving room for improvement. In recent years, FOPLP has emerged as one of its strongest competitors [5].

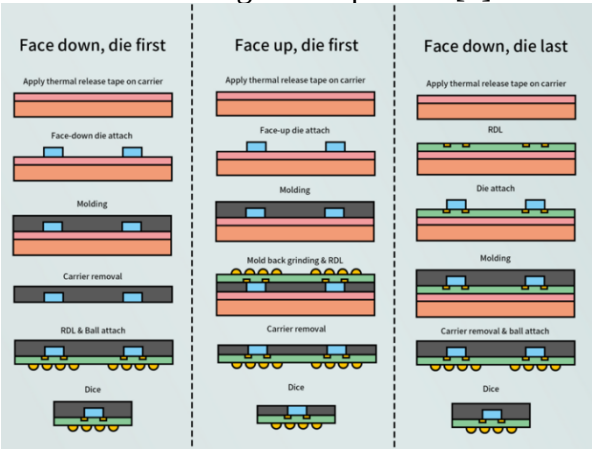


Fig.4 Three types of productions of FOWLP [17]

Fig. 4 clearly illustrates three different FOWLP methods: "Face-down, die-first," "Face-up, die-first," and "Face-down, die-last." Regardless of the method chosen, the core principle remains consistent: first, fix the chips to a substrate using thermal release tape, then attach the substrate to a wafer, followed by encapsulation. After redistributed layer (RDL) wiring on the prepared substrate, the reconfigured wafer is cut into individual packaged chips. This method enables the production of high-integration, high-performance electronic components at a relatively low cost.

As mentioned earlier, when chip sizes are smaller, the time required to populate the wafer during reconstruction increases; moreover, at the same precision level, alignment shifts are more likely to occur, which can negatively impact the performance of electronic components. Additionally, processing FOWLP on smaller wafers results in more material waste, driving up costs. In contrast, using FOPLP technology with 600 mm panels increases component production volume. Furthermore, FOPLP offers higher material utilization, which also reduces costs. The thinner profile achieved by FOPLP is also more conducive to 3D stacking technology [18, 19].

Nevertheless, FOPLP also has unique challenges. One of the most frequently cited issues is warpage and deformation caused by a mismatch between the larger panel and the underlying materials in terms of coefficient of thermal expansion (CTE)—this mismatch leads to deviations in lithography and chip alignment. Another key issue is that as panel size increases, even minor errors can result in the entire panel being scrapped, reducing yield rates.

In any case, FOPLP provides a new perspective for solving industry challenges. Many companies have already begun exploring the development of FOPLP technology, and they are likely to achieve more remarkable results in the future.

4. Architecture

Compared to other innovative approaches that have already delivered results in the industry, architectural innovations aim to fundamentally disrupt the current semiconductor sector, making their implementation significantly challenging. Compared to other revolutionary methods discussed in this paper, architectural innovations are still largely in the early stages of research and development. However, given their enormous future potential, the author believes it is necessary to dedicate some space to them in this paper.

4.1. Optical Computing

In the traditional computer industry, electronics have long been taken for granted as the transmission medium. However, as the pursuit of more advanced process technology and higher computational speed intensifies, optical computers have emerged as a more ideal architectural solution.

In traditional electronic computers, binary information is typically represented by the voltage levels of electrical signals. In contrast, optical computers utilize the properties of light—such as intensity, phase, polarization, or frequency—to encode information. This difference in signal transmission logic leads to distinct approaches for implementing logic gates in the two types of computers. In conventional electronic computers, logic gates are realized by interrupting or connecting electrical currents and controlling voltage levels to transmit corresponding binary signals. In optical computers, however, logic gates rely on light interference and nonlinear propagation under specific conditions (the Fig. 5 shows a model of a NOT gate constructed using optical principles) [20].

The advantages of optical computers over traditional electronic computers are evident. First, since light propagates at speeds far exceeding the movement of electrons in components, optical computers exhibit negligible propagation delay. Second, compared to electronic computers—especially those with smaller process scales—the energy required to generate and transmit optical signals is significantly lower than that needed for electronic signals. Combined with more efficient logic gate systems, optical computers consume less energy than their electronic counterparts. Finally, optical computers maintain parallel processing capabilities even at microscopic scales, without the mutual interference inherent in electronic systems (e.g., via wavelength-division multiplexing). Collectively, these advantages make them a compelling new computing solution that must be considered in the post-Moore era [21].

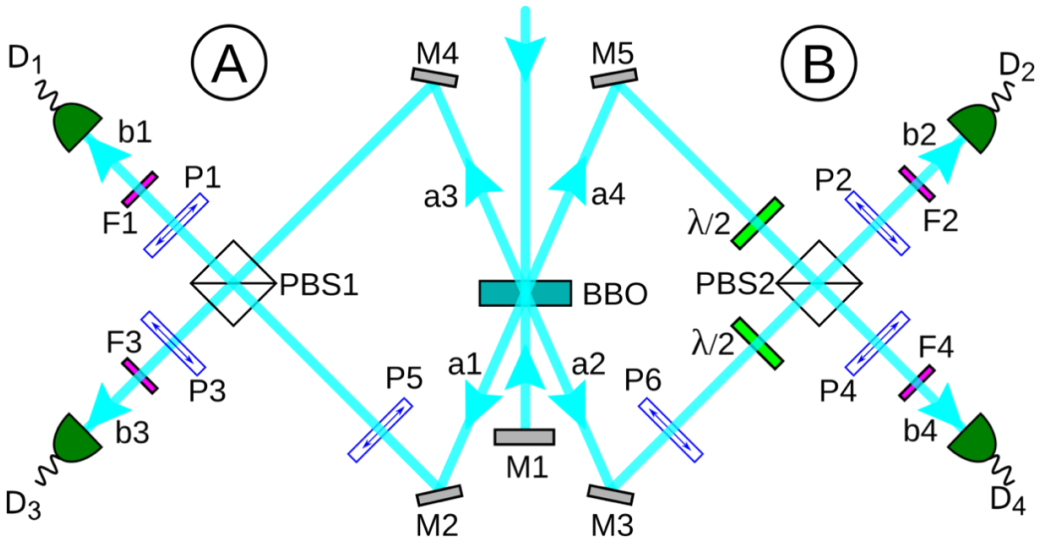


Fig.5 Optical NOT gate [22]

Optical computers do face certain limitations that restrict their practical application. The most significant issue is that photons cannot be easily stopped or stored like electrons—and this makes practical optical memory extremely challenging to develop. Furthermore, while photons avoid collisions with lattice atoms and the leakage currents that plague electrons, they scatter due to waveguide surface roughness, leading to signal attenuation. Consequently, optical computing remains suitable only for applications like Graphics Processing Units (GPUs) (which require massive parallel processing), while optical computing remains inadequate for Central Processing Units CPUs— which demand complex and highly integrated processing. Nevertheless, the industry can still anticipate that computer technologies with high parallelism, high speed, and low power consumption will make their mark in the GPU field.

4.2. Neuromorphic Computing

Another prominent innovative architecture is neuromorphic computing, which fundamentally challenges the traditional von Neumann architecture by mimicking the operating principles of the human brain. Its current applications are typically closely associated with spiking neural networks (SNNs)—networks that simulate the spiking behavior of biological neurons, as shown in Fig. 6.

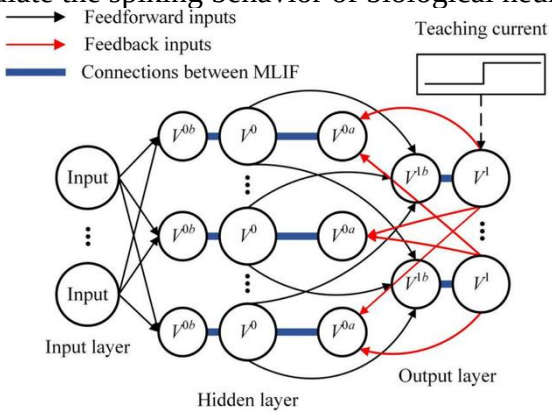


Fig.6 Architecture of spiking neural network [23]

In traditional computer architectures, the processor and memory units are separated and connected via a bus. The control unit first fetches an instruction segment from memory, decodes it, and finally sends data to the corresponding arithmetic unit. In contrast, neuromorphic computing based on SNNs adopts a fundamentally different approach. By mimicking the human brain, this architecture eliminates the spatial separation between the processor and memory, instead consisting of a series of artificial neurons. Each neuron possesses an internal property called Membrane Potential (MP), which represents the cumulative amount of information received by that neuron. Information is transmitted to other neurons via a spike only when the MP reaches a certain threshold. After firing, the MP resets to a lower value. When no new spikes are input, the neuron's MP gradually decays [24]. With this architecture, eliminating the traditional spatial separation between the processor and memory enables significantly faster information transmission. Furthermore, the unique threshold mechanism reduces substantial energy dissipation, resulting in a relatively ideal overall power consumption level.

However, it also faces challenges. At the information level, converting continuous data into discrete spikes carries the risk of information loss. Additionally, the lack of a mature software and hardware ecosystem means building a neuromorphic computing system requires developing specialized hardware and algorithms almost entirely from scratch [25].

5. Materials

5.1. Two-Dimensional Materials

In the innovation process of the traditional semiconductor industry, two-dimensional (2D) materials—with graphene as the representative—are gradually attracting increasing attention. Compared to the more radical approaches mentioned earlier, 2D materials represent a more feasible solution and require lower research and development costs. Fig. 7 shows graphene in photocatalytic applications.

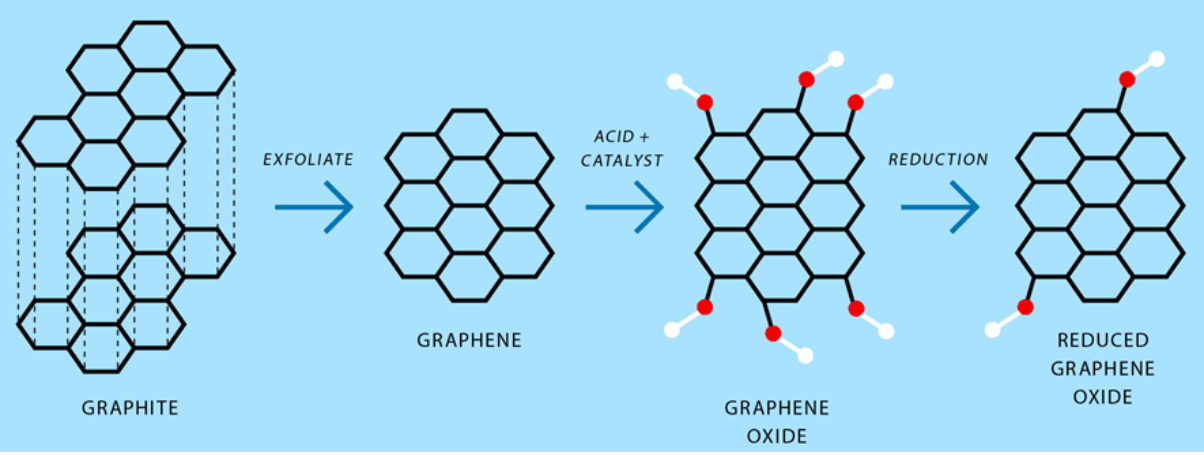


Fig.7 Graphene for Photocatalytic Applicants [26]

Their unique advantages can be observed by comparing their microstructure with that of traditional metallic wires. In conventional wires, atoms form an organized, regular lattice structure. When metals form a bulk structure, their structure forms an electron sea consisting of freely moving electrons. These mobile electrons enable current to flow in the direction of an electric field. However, during their random motion, electrons inevitably collide with vibrating atoms—this increases resistance and power consumption. In two-dimensional materials like graphene, the number of atomic layers is drastically reduced. This minimizes collisions between electrons and the atomic lattice, thus leading to lower resistance and reduced power consumption [27].

5.2. Topological Insulators

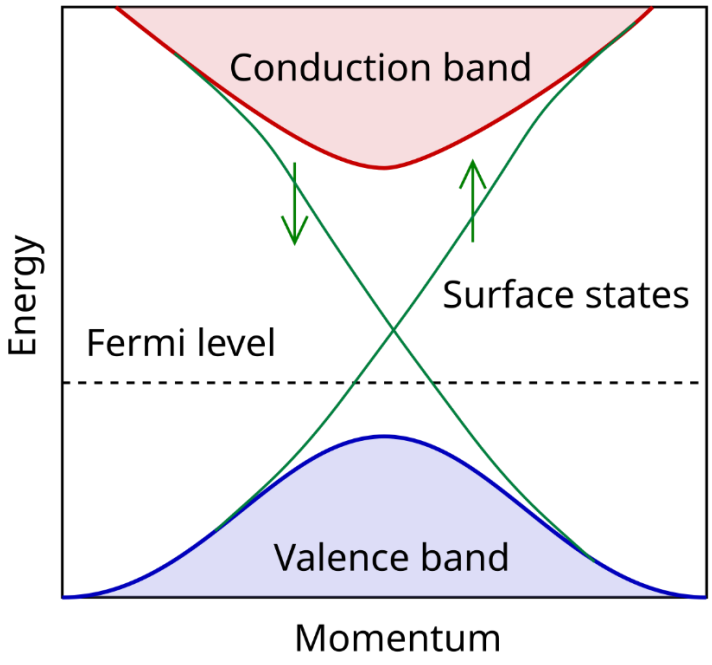


Fig.8 Topological insulator [28]

Another imaginative solution involves combining spintronics with topological insulators to create electronic devices capable of near-lossless high-speed data transmission, as shown in Fig. 8. Their advantages can be elucidated by understanding their respective operating principles.

Spintronics utilizes the spin of electrons to transmit information, requiring less energy than traditional electronics that rely on moving electric charges. Since electron spin does not depend on electric current to maintain itself, information can remain stored even when power is disconnected. Finally, by eliminating most of the wiring required for electron transfer, spintronics enables highly integrated electronic components [29].

Topological insulators are unique materials: they are insulating internally yet conductive on their surfaces. Their distinctive topological properties allow electrons to move extremely fast while preserving the spin characteristics of electrons, thus providing a fundamental platform for spintronics.

6. Conclusion

In summary, Moore's Law—which has long guided the semiconductor industry—is on the verge of being breached. The sector now faces dual challenges of physical limitations and economic viability, forcing the sector into the post-Moore era.

Despite the challenges facing the traditional silicon-based semiconductor chip industry, all these solutions collectively shape a diverse future for the industry. The advent of the post-Moore era is not merely a crisis for the traditional semiconductor chip industry that has developed over the past few decades, but also an opportunity for the industry to re-examine itself, break free from a singular evolutionary path, and collaboratively explore future possibilities across multiple dimensions.

This paper explores potential solutions from multiple angles, including advanced packaging techniques, innovative system architectures, and emerging materials technologies. It aims to contribute to overcoming current obstacles and chart a path forward for the industry.

References

[1] Mark S. Lundstrom, Muhammad A. Alam, Moore’s law: The journey ahead, High-performance electronics will focus on increasing the rate of computation, Science, 2022, 378(6621): 722-723.

[2] Lu X. M., Wang X. G. Incorporating heisenberg’s uncertainty principle into quantum multiparameter estimation. Physical Review Letters, 2021, 126(12): 120503.

[3] Kumar S. Fundamental Limits to Moore's Law, Department of electrical engineering, submitted as coursework for PH250, Stanford University, 2012.

[4] Mirzaie M., Hojbota C. I., Kim D. Y., et al. All-optical nonlinear Compton scattering performed with a multi-petawatt laser. Nature Photonics, 2024, 18(11): 1201-1205.

[5] Lee T. C., Chang Y., Hung C. P., et al. Advanced Packaging from FOWLP to FOPLP: Development of fan-out chip last in 300 Mm panel. 2025 IEEE 75th Electronic Components and Technology Conference. IEEE, 2025.

[6] McMahon P. L. The physics of optical computing. Nature Reviews Physics, 2023, 5(12): 717-734.

[7] Rathi N., Chakraborty I., Kosta A., et al. Exploring neuromorphic computing based on spiking neural networks: algorithms to hardware. ACM Computing Surveys, 2023, 55(12): 1-49.

[8] Fan S., Cao R., Wang L., et al. Quantum tunneling in two - dimensional van der Waals heterostructures and devices. Science China Materials, 2021, 64 (10): 2359-2387.

[9] Krishnamoorthy H. N. S., Dubrovkin A. M., Adamo G., et al. Topological insulator metamaterials. Chemical Reviews, 2023, 123(8): 4416-4457.

[10] Venturelli D., Hamze M., Somma R. D., et al. Computational multiqubit tunnelling in programmable quantum annealers. Nature Communications, 2016, 7: 10327.

[11] Sánchez-Trujillo D. J., Osorio-Maldonado L. V., Prías - Barragán J. J. emperature dependence of electrical conductivity and variable hopping range mechanism on graphene oxide films. Scientific Reports, 2023, 13(1): 4810.

[12] Compton effect [EB/OL]. Encyclopædia Britannica, 2025-09-03.

[13] Khalaf M., Kaminer I. Compton scattering driven by intense quantum light. Science Advances, 2023, 9(1): eade0932.

[14] Law of diminishing marginal utility. Economics, Youtube, 2025.

[15] Li X., Hsee C. K., Inman J. J., Plassmann H. The psychology of marginal utility. Journal of Consumer Research, 2020, 47(4): 475-499.

[16] Kim C., Smith A. N., Kim J., et al. Outside good utility and substitution patterns in direct utility models. Science Direct, 2023.

[17] Xu H., Huang X., Zhuang Z., et al. Hierarchical Partitioning-Based Inter-Chip Redistribution Layer Routing for Fan-Out Wafer-Level Packaging. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2025.

[18] Kim D. H., Lee J. E., Choi G., et al. Study of reliable via structure for Fan Out Panel Level Package (FoPLP). 2022 IEEE 72nd Electronic Components and Technology Conference. IEEE, 2022: 1063-1069.

[19] Lim J. A., Park Y. M., De Vera E., Kim B. C., Dunlap B. 600mm Fan-Out Panel Level Package (FOPLP) as a scale up alternative to 300mm Fan-Out Panel Level Package (FOWLP) with 6-sided die protection. 2021 IEEE 71st Electronic Components and Technology Conference. IEEE, 2021: 1017-1022.

[20] Zhu H. H., Zou J., Zhang H., Shi Y. Z., Luo S. B., Wang N., et al. Space-efficient optical computing with an integrated chip diffractive neural network. Nature Communications, 2022, 13(1): 1-9.

[21] McMahon P. L. The physics of optical computing. Nature Reviews Physics, 2023, 5(12): 717-734.

[22] Psaltis D. Programming nonlinear propagation for efficient optical learning machines[J]. Advanced Photonics, 2024, 6(1).

[23] Tan C., Šarlija M., Kasabov N. Spiking Neural Networks: Background, Recent Development and the NeuCube Architecture. Neural Processing Letters, 2020, 52(2): 1675-1701.

[24] Rathi N., Chakraborty I., Kosta A., Sengupta A., Ankit A., Panda P., Roy K. Exploring neuromorphic computing based on spiking neural networks: algorithms to hardware. ACM Computing Surveys, 2023, 55(12): 1-49.

[25] Kudithipudi D., Schuman C., Vineyard C. M., Pandit T., Merkel C., et al. Neuromorphic computing at scale. Nature, 2025, 637(8047): 801-812.

[26] Ahmed A. Mechanical Characterization of Graphene Nanoparticles. Smart Nanomaterials Technology, 2023.

[27] Das S., Sebastian A., Pop E., McClellan C. J., Franklin A. D., Grasser T., et al. Transistors based on two - dimensional materials for future integrated circuits. Nature Electronics, 2021, 4 (11): 786-799.

[28] Song Y. L., Shang S. X., et al. Disorder-induced phase transitions in a two-dimensional magnetic topological insulator system. Journal of Physics: Condensed Matter, 2025, 37(9).

[29] Zhang Y., Feng X., Zheng Z., Zhang Z., Lin K., et al. Ferrimagnets for spintronic devices: from materials to applications. Applied Physics Reviews, 2023, 10(1): 011301.