Applied Engineering, Innovation, and Technology

Review Article

Future Directions in Semiconductor Processing: Scaling, Integration, and the Sustainability Imperative

Ramatu Al-hassan[1], Edmund Dasori Azundow[2], Yaw Amankrah Sam-Okyere[3,*], Emmanuel Osei-Kwame[3], Nii Ayitey Freddie Aryee[3]

[1]Department of Professional Science, Middle Tennessee State University, Murfreesboro, Tennessee, United States
[2]Department of Electrical Engineering, Eastern Illinois University, Charleston, Illinois, United States
[3]Department of Electronics Engineering, Norfolk State University, Norfolk, Virginia, United States
*Correspondence: y.a.sam-okyere@spartans.nsu.edu

https://doi.org/10.62777/aeit.v3i1.97

Received: 6 October 2025
Revised: 21 January 2026
Accepted: 14 April 2026
Published: 30 May 2026

Abstract: The global semiconductor industry has navigated a period of intense innovation and systemic challenges between 2020 and 2025. Driven by the exponential demands of Artificial Intelligence (AI), 5G/6G communication, and high-performance computing (HPC), the sector has pursued a dual strategy of continued transistor scaling and sophisticated heterogeneous integration. This review systematically analyzes the critical advancements and challenges within this period. We detail the fundamental architectural shift from FinFET to Gate-All-Around (GAA) transistors, enabling the 3-nanometer (nm) and 2-nm nodes, and the adoption of Extreme Ultraviolet (EUV) lithography for High-Volume Manufacturing (HVM). Concurrently, advanced packaging techniques, such as hybrid bonding and the standardization of chiplet architectures via the Universal Chiplet Interconnect Express (UCIe), have emerged as indispensable means to circumvent planar scaling limits. Economically, the industry has contended with escalating capital expenditure (CapEx) and the severe global chip shortage (2020–2023), prompting widespread government intervention, notably through the U.S. CHIPS Act and the EU Chips Act. Crucially, the review addresses the intensifying sustainability mandate, examining the challenges posed by high-Global Warming Potential (GWP) gas emissions, soaring water consumption, and the necessary transition toward circular economy principles within the fabrication environment. The findings underscore that future progress is contingent upon balancing relentless performance demands with resilient supply chains and comprehensive environmental stewardship.

Keywords: semiconductor manufacturing, gate-all-around, high-NA EUV lithography, chiplets.

1. Introduction

The semiconductor industry forms the indispensable foundation of the modern digital economy, powering applications ranging from consumer electronics and automotive systems to national infrastructure and the ongoing AI revolution. Propelled

by surging demand, the global semiconductor market reached approximately USD 573 billion in 2022 [1]. For decades, technological progress has been dictated by Moore's Law, the consistent doubling of transistor density, which traditionally focused on dimensional scaling of planar structures [2]. However, as feature sizes approached atomic limits in the early 2020s, the traditional paradigm faced insurmountable physical constraints, including severe short-channel effects (SCEs), quantum tunneling, and power density challenges [3]. The last 5 years have marked a definitive pivot for the industry, characterized by two primary, co-dependent strategic thrusts:

i. Architectural Scaling: Implementing revolutionary device structures (Gate-All-Around) and advanced patterning technologies (EUV and High-NA EUV) to push scaling past the 5-nm node [4];
ii. System-Level Integration: Leveraging heterogeneous integration and advanced packaging to deliver performance and efficiency gains when transistor scaling slows down [5].

Simultaneously, the industry faced unprecedented external pressures. The global chip shortage (2020–2023), instigated by the COVID-19 pandemic, geopolitical tensions, and climate events such as the Taiwan drought, exposed critical fragilities in global supply chains [6]. This confluence of technological difficulty and systemic risk spurred massive government intervention and placed an urgent spotlight on the environmental footprint of highly energy- and resource-intensive fabrication processes. This review paper provides a focused analysis of the semiconductor processing landscape from 2020 to 2025. Section 2 examines core logic scaling technologies (GAA and EUV). Section 3 explores the "More than Moore" strategy of heterogeneous integration. Section 4 discusses the dominant market and architectural drivers, including AI and emerging post-CMOS concepts. Section 5 analyzes the critical economic and geopolitical dynamics. Finally, Section 6 addresses the intensifying environmental and sustainability challenges confronting the sector.

2. Device Architecture and Patterning

The continuation of Moore's Law in the 2020–2025 period depended fundamentally on innovations in both transistor architecture and lithography.

2.1. Transition to Gate-All-Around Transistors

By the 5-nm node, the industry-standard Fin Field-Effect Transistor (FinFET), which succeeded the planar MOSFET, began to reach its theoretical limits, particularly in maintaining electrostatic control over the channel [7]. To sustain performance and efficiency gains required by AI and mobile platforms, leading foundries initiated the transition to the Gate-All-Around (GAA) transistor architecture. GAA transistors, typically employing a nanosheet or nanowire structure, offer superior electrostatic control by fully wrapping the gate around the channel. This allows for better leakage current management and enables operation at lower supply voltages, significantly improving the power-delay product (PDP). The architectural shift became a commercial reality in this period:

- 3-nm Node Adoption: Samsung was the first to announce the production of its 3-nm process (known as 3GAE) in 2022, utilizing its proprietary Multi-Bridge-Channel FET (MBCFET™) nanosheet architecture [8]. Compared to its preceding 5-nm FinFET process, Samsung reported up to a 45% reduction in power consumption or a 23% improvement in performance. TSMC also introduced its 3-

nm process (N3) during this period, claiming a 25–30% power reduction or a 10–15% speed increase over its 5-nm node.

- 2-nm and Beyond: The competitive roadmap extended immediately to the 2-nm node, which is slated for volume production around 2025. IBM demonstrated a 2-nm nanosheet chip prototype in 2021, showcasing the potential for substantial gains [9]. TSMC's 2-nm process, which will also utilize stacked nanosheet transistors, aims to deliver a 15% speed uplift or 30% power saving compared to its 3-nm node [10].

Figure 1 presents a side-by-side comparison of a FinFET and a gate-all-around (GAA) nanosheet FET. (a) key components, including shallow trench isolation (STI), source/drain (S/D) epitaxy, and a high-k metal tri-gate. (b) The GAA nanosheet FET is shown with STI, S/D epitaxy, bottom dielectric isolation (BDI), and a high-k metal gate that fully surrounds the channel. Certain features, such as BDI and the isolation between the gate and S/D, are exclusive to the GAA nanosheet FET architecture [11].

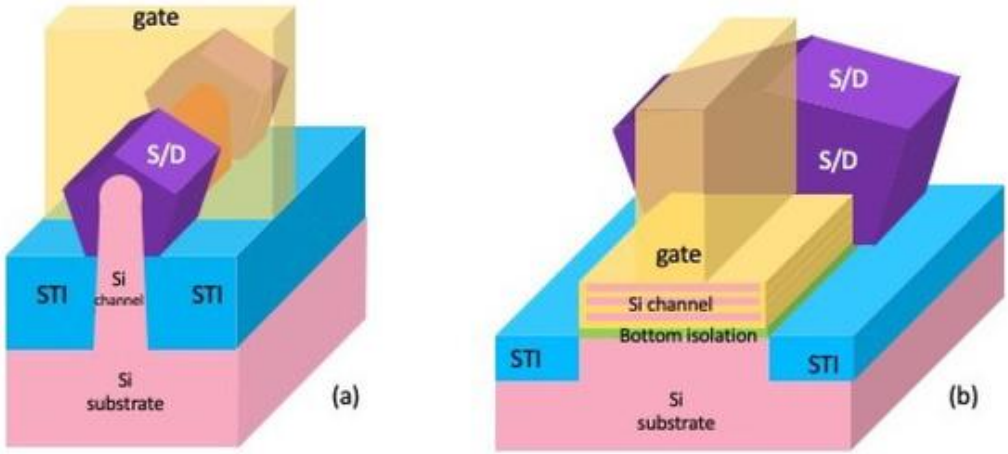

Figure 1. Side-by-Side Comparison of: (a) FinFET and (b) Gate-all-around (GAA) nanosheet FET.

Transitioning to gate-all-around architectures from tri-gate designs becomes necessary as FinFET scaling beyond the 7 nm node intensifies short-channel effects (SCEs). Within the realm of gate-all-around structures, the semiconductor industry has explored options like nanowires, which offer superior electrostatic control, and nanosheets, which deliver higher "on" current and better electrostatic performance compared to FinFETs. Figure 1 shows a FinFET alongside a gate-all-around nanosheet FET, highlighting their main components. Shared elements between the two include shallow trench isolation, source/drain epitaxies, and high-k metal gates, while their structural distinctions lie in the tri-gate configuration of FinFETs versus the all-encompassing gate of nanosheets. For enhanced performance, nanosheets are stacked vertically, in contrast to FinFETs, where a single fin forms the device. This transition introduces significant fabrication challenges, including the precise epitaxial growth of Silicon Germanium (SiGe)/Silicon (Si) superlattice structures at temperatures below 650°C and the subsequent highly selective removal of the SiGe sacrificial layers to form the nanosheet channels. Furthermore, achieving identical transistor behavior critical for analog design (matching) is increasingly difficult due to the acute sensitivity of GAA structures to line edge roughness (LER) and stress effects. Performance metrics for 3-nm GAA implementations is shown in Table 1.

Table 1. Reported and independently estimated performance metrics for 3-nm GAA implementations.

Foundry	Process	Transistor Type	Power Reduction vs 5nm	Performance Gain vs 5nm	Area Scaling	Yield/Maturity (Reported)
Samsung	3GAE	Nanosheet GAA (MBCFET)	45% (claimed)	23% (claimed)	~16%	Early HVM, limited customer adoption
TSMC	N3	FinFET (first-gen), GAA at N2	25-30% (claimed)	10-15% (claimed)	~1.6 × density	Higher yield, broad customer uptake
IBM	2 nm demo	Nanosheet GAA	~45% (lab data)	~45%	Research only	Not applicable

From a device physics perspective, GAA architectures demonstrate measurable improvements in short-channel effects compared with FinFETs at equivalent gate lengths. Experimental and simulation studies report drain-induced barrier lowering (DIBL) reductions of approximately 30–50% and subthreshold swing improvements approaching the thermal limit (~65–70 mV/dec) for nanosheet GAA devices, compared to typical FinFET values exceeding 80 mV/dec at similar nodes. These gains directly translate into reduced leakage currents and enable lower operating voltages, reinforcing the suitability of GAA transistors for energy-constrained AI and mobile applications. While Samsung achieved an important milestone by introducing GAA transistors earlier at the 3-nm node, this first-mover advantage came with notable trade-offs in manufacturing maturity.

Early reports indicate that Samsung's 3GAE process experienced yield challenges typical of first-generation GAA integration, limiting initial commercial adoption. In contrast, TSMC delayed full GAA deployment until its 2-nm node, prioritizing incremental optimization of FinFET-based N3 for yield stability and ecosystem readiness. This strategic divergence highlights a fundamental trade-off between architectural leadership and process maturity. Samsung accelerated electrostatic control benefits at the cost of yield learning, whereas TSMC favored a more conservative transition that improved manufacturability and customer confidence. Such differences underscore that commercial viability at advanced nodes is driven as much by yield and defect density as by transistor-level performance metrics.

2.2. Sustaining Scaling with Extreme Ultraviolet Lithography

Continued scaling for 7-nm, 5-nm, and 3-nm nodes depended heavily on the maturity and large-scale deployment of Extreme Ultraviolet (EUV) lithography, utilizing 13.5-nm light [12]. Since its initial deployment for High-Volume Manufacturing (HVM) around 2019, EUV successfully replaced the complex, multi-patterning 193-nm immersion lithography steps for the most critical layers. The use of EUV was central to enabling TSMC's 5-nm, 4-nm, and 3-nm nodes [13]. As the 3-nm node approached resolution limits, the focus shifted to the next generation: High-Numerical-Aperture (High-NA) EUV lithography. Figure 2 summarizes how UCIe underpins heterogeneous systems that increasingly co-evolve with patterning limits. Panel (a) depicts an open, multi-vendor chiplet platform where logic, I/O, and accelerators are mixed on one package; this modularity reduces monolithic die size pressure as lithography reaches stochastic limits. Panel (b) situates UCIe across 2D and 2.5D options (standard bumps, micro-bumps, EMIB, CoWoS, fan-out), while panel (c) stacks the PHY/protocol/software layers that make chiplet interoperability practical. Panel (d) shows multi-module configurations (e.g., PCIe/CXL fabrics) that scale system bandwidth without relying solely on further pitch shrink [14].

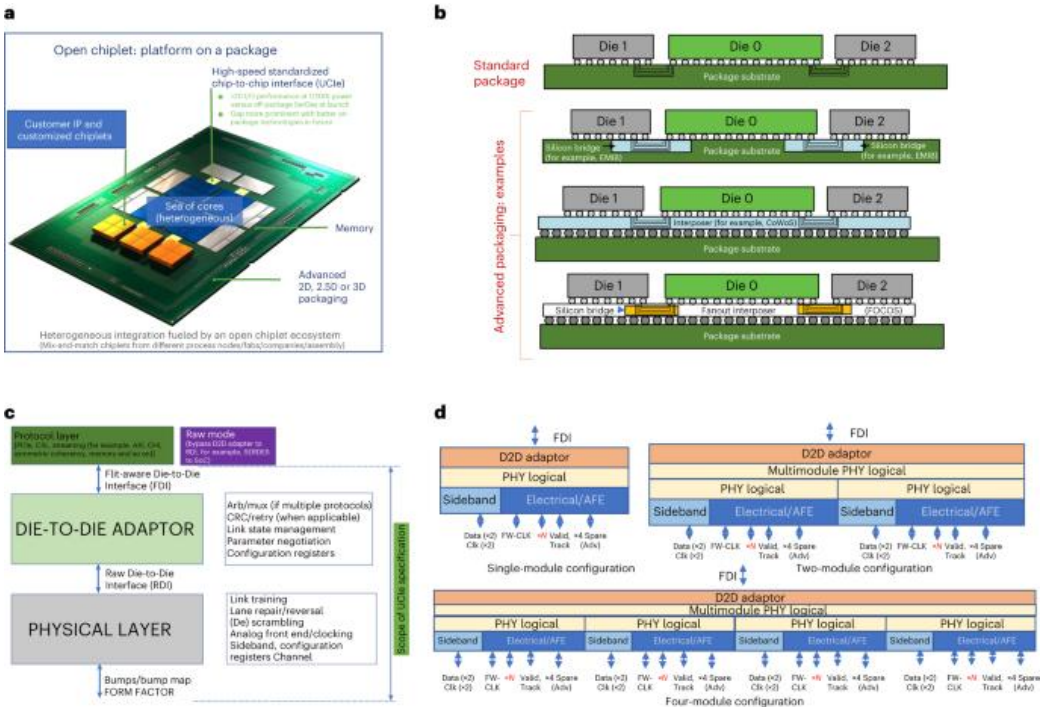

Figure 2. A Heterogeneous Open Chiplet On-Package System: (a) an open, multi-vendor chiplet platform on one package, (b) UCIe across 2D and 2.5D configurations, (c) layered structure in the UCIe, and (d) multi-module setups.

Although High-NA EUV offers substantial resolution benefits, its economic implications are nontrivial. The reduced imaging field and lower throughput relative to 0.33-NA systems increase cost per wafer, while the capital cost of High-NA tools exceeds USD 300 million per scanner. Consequently, High-NA EUV adoption is expected to be selective and initially confined to the most critical layers at the 2-nm node. This reinforces the need for careful cost–performance co-optimization rather than wholesale replacement of existing EUV infrastructure. Table 2 summarizes the comparison of low-NA and high-NA EUV lithography systems.

Table 2. Comparison of low-NA and high-NA EUV lithography systems.

Parameter	Low-NA EUV	High-NA EUV
Numerical Aperture	0.33	0.55
Typical Throughput	~160-180 wafers/hour	~120-140 wafers/hour
Tool Cost (est.)	~USD 180-200 million	> USD 300 million
Cost per Wafer	Lower	Significantly higher
Primary Node	5nm-3nm	2nm and beyond
Key Limitation	Resolution ceiling	Field size, stitching

In late 2023, the first High-NA EUV tool (the ASML EXE:5000) was shipped, featuring a numerical aperture (NA) of 0.55 [15]. This represents a significant advancement over the previous 0.33-NA systems, promising a 1.7× shrink in feature size and up to a 2.9× increase in transistor density. High-NA EUV is positioned as the core technology to enable the 2-nm node and beyond, with the ecosystem developing rapidly to support its HVM capability by 2025 [16]. High-NA EUV presents new complexities. Its anamorphic magnification means the image field is smaller than previous EUV scanners, necessitating a technique called "stitching" to expose dies larger than the imaging field [17], [18]. This involves precisely aligning two mask fields at a critical boundary, posing major challenges for design and process control. Furthermore, material optimization, including the co-optimization of chemically amplified resists (CAR) and metal oxide resists (MOR), is

essential to mitigate stochastic variations and maintain critical dimension (CD) stability at sub-3-nm pitches.

3. Architecting Performance

As the cost and complexity of node scaling intensified, the industry concurrently expanded its reliance on Heterogeneous Integration (HI) and advanced packaging to drive system performance, a strategy often termed "More than Moore". HI involves combining multiple dies, fabricated on different optimal process nodes, into a single, high-performance package.

3.1. The Chiplet Ecosystem and Interface Standardization

The concept of the chiplet, a modular die optimized for a specific function, gained significant commercial traction between 2020 and 2025. Chiplets offer several advantages over monolithic designs, including improved yield (as smaller dies have higher yield rates), enhanced design flexibility, and the reuse of optimized Intellectual Property (IP) [19]. Key commercial examples include AMD's latest CPUs and GPUs and Intel's use of chiplet architectures in products like the Stratix 10 FPGAs. A critical development for the broad adoption of chiplets was the effort toward industry standardization of the die-to-die communication interface. The Universal Chiplet Interconnect Express (UCIe) standard, which released version 2.0 in August 2024, was established to provide a ubiquitous interconnect layer at the package level. By leveraging established industry protocols like PCI Express (PCIe) and Compute Express Link (CXL), UCIe aims to simplify the integration of chiplets from different vendors, moving the industry toward a standardized, modular design methodology [20].

Despite its promise, UCIe adoption introduces several practical challenges. Die-to-die interconnects incur higher power overhead and latency compared to on-die wiring, especially for fine-grained coherence traffic. Thermal and mechanical mismatches between heterogeneous chiplets complicate reliability, while multi-vendor integration raises nontrivial issues related to IP qualification, test methodologies, and yield accountability. As a result, most successful UCIe demonstrations to date remain confined to controlled ecosystems rather than fully open, multi-foundry deployments. These constraints indicate that chiplet standardization, while transformative, is not a panacea and must be evaluated within system-level power, latency, and reliability budgets.

3.2. 3D Stacking and Hybrid Bonding

To meet the high-bandwidth requirements of AI and HPC accelerators, which suffer acutely from the "memory wall," vertical integration through 3D stacking became essential [21]. This period saw the maturity and increasing adoption of Hybrid Bonding as the definitive technology for next-generation 3D integration. Traditional vertical connections, such as micro-bumps and Through-Silicon Vias (TSVs), face yield and reliability challenges as interconnect pitches shrink below 10 micrometers (µm). Hybrid bonding bypasses these limitations by directly bonding the metal pads and dielectric layers of two wafers (wafer-to-wafer, W2W) or a die to a wafer (die-to-wafer, D2W) [22]. This process achieves extremely high interconnect density, enabling pitches below 10 µm and even toward the sub-micron scale, making it the only viable solution for ultra-dense 3D applications. Hybrid bonding is the enabling technology for sophisticated High-Bandwidth Memory (HBM) stacks (ranging from 8-high to 24-high) and sophisticated logic-on-logic or logic-on-memory integration platforms, such as those offered by TSMC and Intel. While offering superior performance, 3D stacking presents significant manufacturing and operational challenges:

- Alignment and Precision: Maintaining sub-micron alignment precision during the bonding process is exceptionally difficult, as misalignment can cause performance degradation and defects.
- Thermal Management: The dense stacking of active layers severely complicates power delivery and thermal dissipation. Localized overheating must be actively suppressed through optimized material composition, metal density, and dedicated thermal management solutions to ensure the reliability of high-layer HBM stacks.

Figure 3 outlines hybrid bonding, which directly bonds dielectric-to-dielectric and Cu-to-Cu interfaces, enabling interconnect pitches from <10 µm toward the sub-micron regime. The top sequence shows the flow planarize and align, form dielectric contact, then achieve Cu–Cu bonding, followed by low-temperature anneal, while the cross-section emphasizes face-to-face copper pads with minimal parasitic. Compared with micro-bumps/TSVs, hybrid bonding delivers vastly higher I/O density and lower resistance/inductance, which is decisive for logic-on-memory and SRAM-on-logic stacks. The practical constraints are sub-µm alignment, surface cleanliness/planarity, and thermal budgets that protect device characteristics during post-bond steps. Quantitatively, hybrid bonding enables interconnect densities exceeding 10^6 I/O per mm², compared to approximately 10^4–10^5 I/O per mm² achievable with advanced micro-bump technology. Electrical benefits include significantly reduced parasitic resistance and inductance, while thermal resistance across bonded interfaces can be reduced by 20–30% due to the elimination of solder interfaces. These characteristics make hybrid bonding indispensable for bandwidth-intensive and thermally constrained applications such as logic-on-HBM integration in AI accelerators.

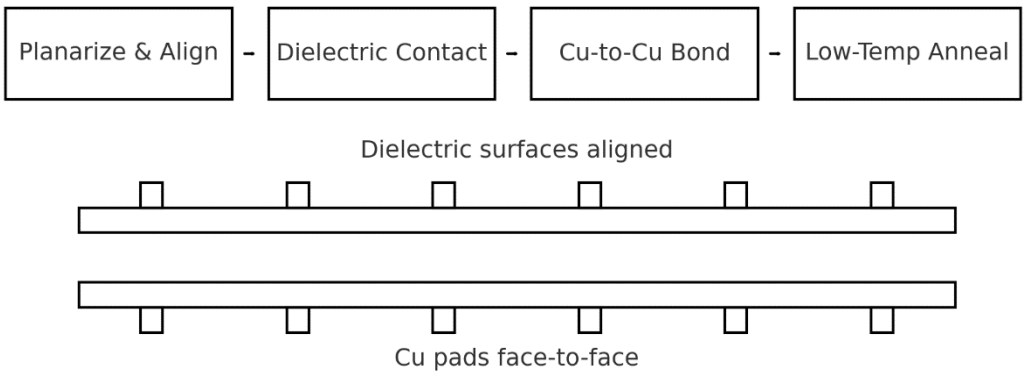

Figure 3. Hybrid bonding (dielectric-to-dielectric + copper-to-copper).

4. Market Drivers and Post-CMOS Directions

The pace and direction of semiconductor development were profoundly shaped by the demands of emerging computing paradigms, particularly AI.

4.1. Domain-Specific Processors and AI Acceleration

The exponential growth in deep learning and machine learning workloads cemented Artificial Intelligence (AI) as the singular most powerful driver for cutting-edge chip technology. AI and HPC were identified as the primary adopters of the 3-nm node by 2025, prioritizing the maximum performance and efficiency that advanced nodes provide. This demand fueled the proliferation of domain-specific processors, specialized accelerators designed for massive parallelism, including GPUs, Tensor Processing Units (TPUs), and dedicated ASICs. A defining feature of these high-performance chips is the direct integration of High-Bandwidth Memory (HBM) on the package, utilizing 3D stacking

via hybrid bonding, to alleviate the severe memory bandwidth bottlenecks inherent in large AI models [23]. Research into neuromorphic and in-memory computing remained an active area. These architectures aim to achieve massive energy efficiency gains by emulating neural networks directly in hardware, using novel devices (such as memristors or phase-change memory) to combine processing and memory functions.

Figure 4 illustrates a conventional interposer channel for HBM, focusing on the signal path characteristics that dominate crosstalk and insertion loss. The drawing highlights the close spacing of parallel channels, the return path geometry, and points where far-end crosstalk (FEXT) becomes problematic at very high data rates. In the context of your AI accelerator discussion, this baseline is important: it motivates channel engineering (e.g., vertical-tabbed vias, shielded lines, or alternative return paths) that reduces FEXT and preserves eye openings as per-pin speeds rise [24].

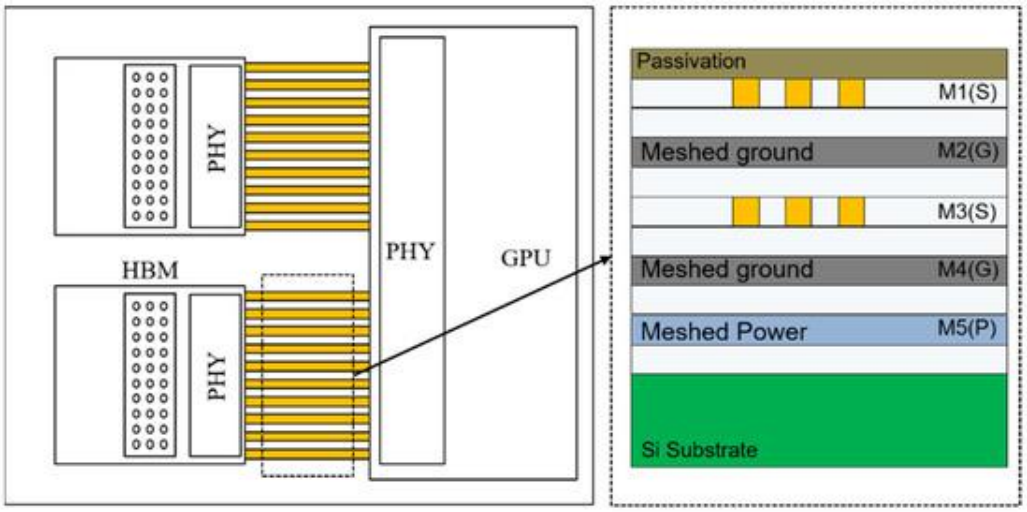

Figure 4. Conventional HBM interposer channel [25].

4.2. Emerging Post-CMOS Paradigms

Looking beyond conventional scaling, the period saw significant investment and milestones in alternative computing technologies that could potentially complement or succeed CMOS in the future.

1. Silicon Photonics (SiP)
   As electrical interconnects struggle with speed, power, and bandwidth limits over distance, Silicon Photonics, using light instead of electrons for data transmission, has moved closer to commercial viability. SiP is becoming critical for high-speed optical I/O in data centers and HPC systems. Gartner included photonic computing in its 2025 Hype Cycle, indicating its increasing traction among industry leaders and its potential for specialized, low-energy, high-throughput AI processing [26].
2. Quantum Computing
   Quantum technology transitioned from pure academic speculation to a major focus of global investment during this period, with China leading in patent filings between 2020 and 2024. Hardware advancements focused on scaling qubit counts and improving coherence times. IBM, for instance, pushed the limits with its *Nighthawk* processor (133 fixed-frequency qubits), while Intel continued to champion silicon spin qubits for their potential compatibility with existing CMOS manufacturing infrastructure. Exploratory research also began to bridge these fields, investigating the concept of neuromorphic quantum computing, utilizing

elements like quantum memristors as artificial synapses in quantum neural networks [27].

5. Economic and Geopolitical Imperatives

Recent years have demonstrated the semiconductor industry's immense geopolitical significance and its vulnerability to external shocks. The continuous technological migration to 3-nm, 2-nm, and High-NA EUV lithography resulted in a sharp escalation of capital expenditure (CapEx) [28]. New leading-edge fabrication facilities now require investments exceeding $20 billion, an economic reality driven largely by the extraordinary cost of advanced toolsets. This CapEx trend continued upward throughout the period, largely sustained by the explosive demand for advanced AI chips. The global semiconductor market is now projected to exceed $1 trillion by 2030, underscoring the enormous financial scale of the sector. Following cascading global disruptions, including geopolitical tensions, climate events, and the pandemic, companies shifted their strategic focus toward a "cost of resilience" mindset. This involves accepting higher operational costs and greater complexity to build redundant, regional supply chains, ensuring agility and mitigating risks associated with single-source reliance.

Figure 5 compares an INV×1 cell's propagation delay (left) and output transition (slew) delay (right) under the same conditions (input slew = 20 ps, output load = 2.88 fF) for three power-delivery/topology options: a 6-track front-side power rail (FS-PR) cell, a 5-track buried power rail (BPR) cell, and a 4-track backside power (BSP) cell. Across both panels, the p-type device (blue) consistently shows a higher delay than the n-type device (orange), reflecting the typical p/n drive asymmetry. Moving from 6T → 5T → 4T compacts the cell and tightens routing; at iso-load, this tends to increase both propagation delay and transition delay, with FS-PR fastest, BPR in the middle, and BSP (4T) the slowest. The result highlights a key co-optimization point: while BPR/BSPDN can improve IR-drop and global power integrity at the chip level, aggressive track-height reduction and device sizing inside the standard cell can raise the intrinsic inverter delay and output slew, especially for the p-device, unless compensated by drive-strength choices or cell architecture tweaks.

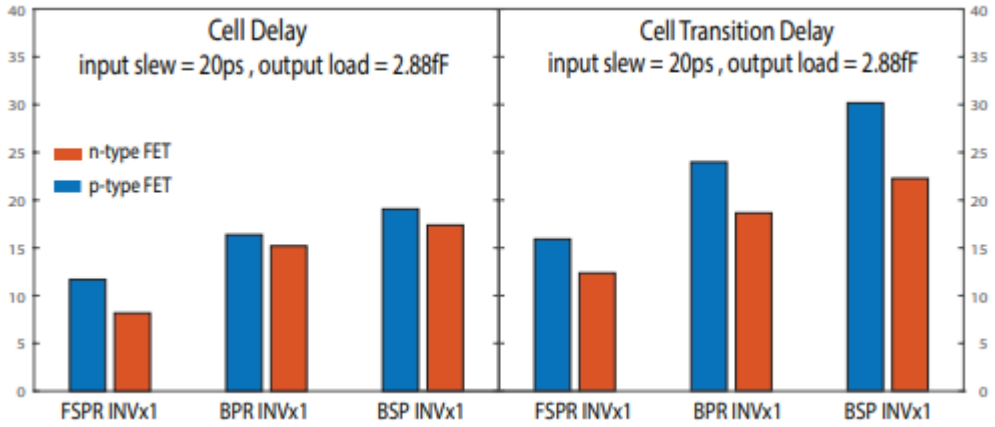

Figure 5. Cell rise and fall delay comparison of 6T FS-PR, 5T BPR, and 4T BSP INVx1 cells [29].

The period was defined by the profound impact of the 2020–2023 global chip shortage. The crisis, exacerbated by canceled orders from the auto industry at the start of the pandemic and compounded by supply disruptions and droughts in key manufacturing regions like Taiwan, severely affected more than 169 industries. For the automotive sector alone, an estimated 9.5 million units of global light-vehicle production were lost in 2021 directly due to the lack of necessary chips. In response, governments

launched massive industrial policy initiatives to bolster domestic production capacity and diversify supply chains:

- U.S. CHIPS Act: The U.S. CHIPS and Science Act (2022) allocated over $52 billion in funding to incentivize domestic semiconductor manufacturing and research and development. Early returns demonstrated optimism, as the Act spurred nearly half a trillion dollars in private sector commitments across the semiconductor ecosystem [30].
- EU Chips Act: The European Union launched its own initiative to increase manufacturing share. By late 2025, semiconductor firms were already calling for a follow-up ("Chips Act 2.0"), highlighting the ongoing need for continuous support to achieve the bloc's ambitious digital goals [31].
- Global Diversification: Other nations also made significant strategic investments. India, for example, approved over $15.2 billion to develop its domestic manufacturing ecosystem. These initiatives collectively underscore the shift of semiconductors from a purely economic commodity to a strategically critical geopolitical resource [32].

6. Environmental Challenges and Mitigation

Semiconductor manufacturing contributes significantly to global greenhouse gas (GHG) emissions across three main categories: Scope 1 (direct emissions from processes), Scope 2 (indirect emissions from purchased electricity), and Scope 3 (upstream supply chain emissions).

Direct emissions (Scope 1) are largely defined by the use of highly potent fluorinated gases (F-gases) such as perfluorocarbons (PFCs), nitrogen trifluoride (NF3), sulfur hexafluoride (SF6), and hydrofluorocarbons (HFCs, including HFC-23) in etching and cleaning steps. These gases possess extremely high Global Warming Potentials (GWPs) [1]. The climate impact of fluorinated process gases is amplified by their exceptionally high Global Warming Potentials (GWPs). For example, nitrogen trifluoride (NF_3), widely used for chamber cleaning, has a GWP of approximately 17,200 times that of CO_2 over a 100-year horizon, while sulfur hexafluoride (SF_6), used in certain etching and insulation processes, has a GWP of approximately 22,800 CO_2-equivalent. Even relatively small leakages or incomplete abatement of these gases can therefore result in disproportionately large climate impacts, underscoring the urgency of effective emission control strategies in advanced fabs.

While direct process emissions are critical, indirect emissions from energy consumption remain the largest source overall. For example, in Taiwan's semiconductor sector in 2020, Scope 2 emissions accounted for approximately 50% of the total, with the complex Scope 3 supply chain emissions contributing about 37%. Efforts across the industry focused on two main areas:

- Abatement: Installing combustion or thermal abatement systems to remove gases like SF6 and other fluorinated greenhouse gases generated during etching. Process optimization is crucial for reducing emissions from persistent gases like NF3, as optimization and dedicated abatement systems are needed to counter the high proportion of gas released into the environment.
- Clean Energy Commitments: Companies made significant public commitments to decarbonize their Scope 2 footprint. For example, STMicroelectronics committed to carbon neutrality for its Scope 1 and 2 emissions (and partially Scope 3) by 2027, with an intermediate goal of sourcing 100% renewable energy by that time [33].

Figure 6 summarizes process-gas abatement for high-GWP fluorinated gases (e.g., NF_3, SF_6, PFCs) used in etch/clean steps. The block diagram traces effluent flow from process tool exhaust through pre-treatment and point-of-use (POU) abatement modules (thermal, plasma, catalytic), with stack monitoring to verify destruction and removal efficiency (DRE). The accompanying data show typical DRE ranges and conditions for stable operation; note that recipes optimized to reduce gas usage often work best when paired with high-efficiency POU systems.

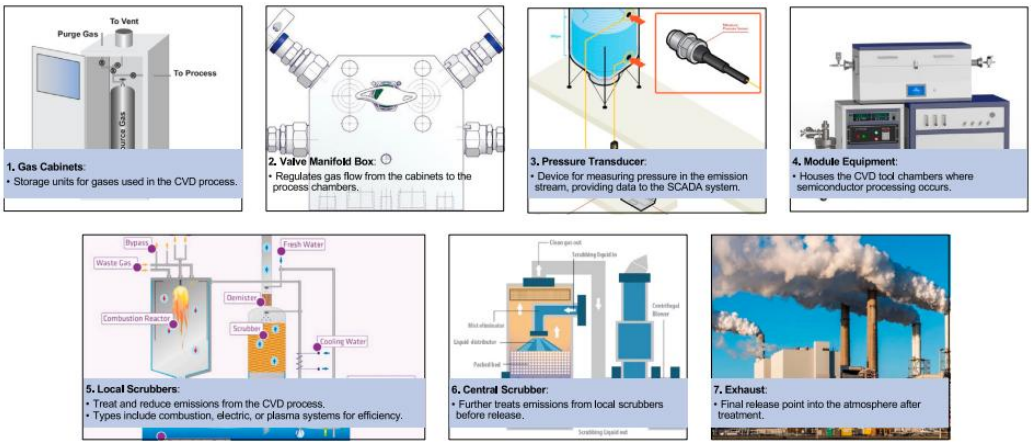

Figure 6. Emission stream diagram for cleaning chemical vapor deposition module tool chambers with pressure transducer [34].

The resource-intensive nature of advanced fabrication amplified existing concerns over water scarcity and hazardous waste management. Semiconductor manufacturing requires vast quantities of ultra-pure water (UPW) for wafer rinsing and cleanroom climate control, with consumption projected to double by 2035. In 2023 alone, a single leading manufacturer consumed a reported 101 million cubic meters (m^3) of water. The fact that many fabs are located in water-stressed regions, such as Taiwan and Arizona, made water management a top strategic risk [35]. Mitigation efforts focused on water reuse, recycling wastewater for use in cooling towers or eventually reprocessing it back into UPW. Leading companies have set aggressive targets, such as the goal of implementing 100% reclaimed water systems. Beyond water, the industry is committed to leveraging the circular economy for materials, focusing on recycling copper and rare elements from waste streams to mitigate reliance on climate-vulnerable extraction processes and high-purity chemical production.

Figure 7 provides an overview of the end-of-life (EOL) cycle for electronic devices, alongside the interconnected water cycle in semiconductor manufacturing, spanning from initial water consumption through wastewater discharge and eventual recycling efforts [36].

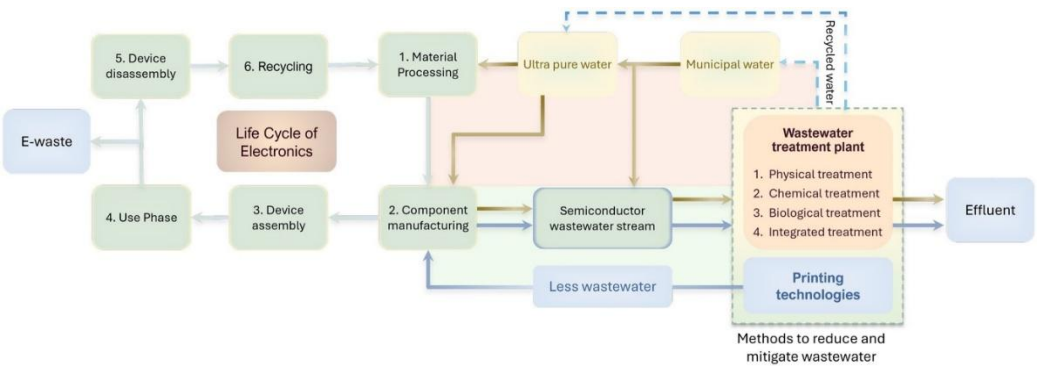

Figure 7. Overview of the end-of-life (EOL) cycle for electronic devices.

Figure 7 maps the water lifecycle in semiconductor manufacturing alongside broader end-of-life (EOL) circularity. On the water side, it traces intake → treatment → ultra-pure water (UPW) → process → reclaim/reuse → discharge, calling out contaminant classes and treatment strategies at each stage. The EOL panel complements this by situating water stewardship within a wider circular-economy perspective—recovery of critical materials, closed-loop chemistries, and design choices that reduce embodied impacts across Scope 3. Achieving near-100% water reclamation presents significant technical challenges, particularly in removing trace contaminants such as silica, heavy metals, and organic residues to meet ultra-pure water (UPW) specifications. Advanced treatment processes are energy-intensive and increase operational costs. As a result, aggressive water recycling targets must be evaluated not only in terms of water conservation but also through a life-cycle energy and carbon lens to avoid unintended trade-offs between water sustainability and greenhouse gas emissions.

Abatement technologies for fluorinated greenhouse gases vary significantly in efficiency, cost, and operational complexity. Thermal abatement systems typically achieve destruction and removal efficiencies (DREs) exceeding 90% but incur high energy consumption and operating costs. Catalytic abatement offers lower operating temperatures and reduced energy demand but is sensitive to catalyst poisoning and requires frequent maintenance. Plasma-based abatement systems provide high DREs for a broad range of gases but introduce additional electrical power consumption and system complexity. Consequently, optimal abatement strategies increasingly involve hybrid configurations tailored to specific process gases, balancing environmental effectiveness with economic feasibility.

7. Conclusions and Future Works

The period from 2020 to 2025 proved to be one of critical transition and profound challenge for the semiconductor industry. The sector effectively countered the limitations of planar scaling through a comprehensive dual strategy: implementing the architectural shift to Gate-All-Around transistors for 3-nm and 2-nm nodes, and aggressively adopting advanced packaging techniques like hybrid bonding to enable high-bandwidth heterogeneous integration. These technological vectors successfully sustained the pace of performance growth, driven largely by the explosive demand for domain-specific AI accelerators. However, this innovation occurred against a backdrop of severe systemic pressures. The global chip shortage laid bare the fragility of highly complex global supply chains, prompting unprecedented government intervention through regional initiatives like the U.S. CHIPS Act and the EU Chips Act. Concurrently, the increasing scale and complexity of fabrication underscored the urgent mandate for environmental sustainability, requiring immediate action on high-GWP gas abatement and resource stewardship, particularly concerning water and chemical recycling. Ultimately, the future viability of the semiconductor industry will depend not solely on its ability to manufacture more advanced chips but on its capacity to do so within a framework that ensures geopolitical resilience, economic sustainability, and responsible environmental impact. The coordinated, multi-stakeholder approach adopted during this period, encompassing design, manufacturing, policy, and environmental accountability, is essential for navigating the next phase of technological evolution.

Looking beyond the 2025 horizon, the trends established in this period suggest that the industry's central challenge will pivot from incremental scaling to systemic, integrated optimization. The continued roadmap for logic scaling points toward the next generation of architectures, beyond nanosheets, such as the implementation of Complementary FET

(CFET) structures that stack N-type and P-type devices vertically. Achieving this will require overcoming even greater challenges in material science, defectivity control, and thermal management within 3D structures. In heterogeneous integration, the focus will be on further standardizing the chiplet ecosystem and pushing hybrid bonding capabilities toward sub-micron pitches, making highly integrated 3D systems the default architecture for performance computing. Design automation, leveraging AI and machine learning, will become critical for managing the vast complexity of co-designing these 3D systems across multiple process technologies.

From an environmental standpoint, the focus must shift from setting targets to full execution. Achieving carbon neutrality will require full decarbonization of Scope 2 power sources, coupled with advanced process-level innovation, including green chemistry and closed-loop systems to eliminate highly persistent F-gases. Furthermore, greater transparency and coordinated action are needed to address the substantial, complex, and often geographically dispersed Scope 3 emissions across the raw material and equipment supply chains. The emergence of post-CMOS paradigms, specifically the industrialization of silicon photonics for high-speed communication and the continued development of error-corrected quantum computing and neuromorphic accelerators, will begin to reshape the foundation of computing itself. While traditional CMOS remains dominant, the integration of these novel technologies into the semiconductor manufacturing ecosystem will define the competitive landscape of the late 2020s and beyond.

References

[1] Y. Yin and Y. Yang, "Sustainable Transition of the Global Semiconductor Industry: Challenges, Strategies, and Future Directions," *Sustainability*, vol. 17, no. 7, p. 3160, Apr. 2025, doi: 10.3390/su17073160.

[2] D. C. Brock, *Understanding Moore's law: four decades of innovation*. Philadelphia: Chemical Heritage Foundation, 2006.

[3] T. N. Theis and H.-S. P. Wong, "The End of Moore's Law: A New Beginning for Information Technology," *Comput. Sci. Eng.*, vol. 19, no. 2, pp. 41–50, Mar. 2017, doi: 10.1109/MCSE.2017.29.

[4] H. H. Radamson *et al.*, "CMOS Scaling for the 5 nm Node and Beyond: Device, Process and Technology," *Nanomaterials*, vol. 14, no. 10, p. 837, May 2024, doi: 10.3390/nano14100837.

[5] M. Manley, A. Victor, H. Park, A. Kaul, M. Kathaperumal, and M. S. Bakir, "Heterogeneous Integration Technologies for Artificial Intelligence Applications," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 10, pp. 89–97, 2024, doi: 10.1109/JXCDC.2024.3484958.

[6] K. N. Ochonogor, G. S. Osho, C. O. Anoka, and O. Ojumu, "The COVID-19 pandemic and supply chain disruption: an analysis of the semiconductor industry's resilience," *International Journal of Technical & Scientific Research Engineering*, vol. 6, no. 1, pp. 7–18, 2023.

[7] C. Hu, "FinFET: The 3D Thin-Body Transistor," in *75th Anniversary of the Transistor*, Wiley, 2023, pp. 135–144. doi: 10.1002/9781394202478.ch12.

[8] Chinese Academy of Cyberspace Studies, "World Information Technology Development," in *World Internet Development Report 2021*, Singapore: Springer Nature, 2023, pp. 69–96. doi: 10.1007/978-981-19-9323-7_3.

[9] S. K. Moore, D. Johnson, M. Harris, E. Waltz, P. Patel, and M. Hampson, "The Latest Developments in Technology, Engineering, and Science: News," *IEEE Spectr.*, vol. 58, no. 8, pp. 5–12, Aug. 2021, doi: 10.1109/MSPEC.2021.9502948.

[10] C. Sheng *et al.*, "Two-Dimensional Semiconductors: From Device Processing to Circuit Integration," *Adv. Funct. Mater.*, vol. 33, no. 50, Dec. 2023, doi: 10.1002/adfm.202304778.

[11] S. Mukesh and J. Zhang, "A Review of the Gate-All-Around Nanosheet FET Process Opportunities," *Electronics (Basel).*, vol. 11, no. 21, p. 3589, Nov. 2022, doi: 10.3390/electronics11213589.

[12] Z. Cui, "Optical Lithography," in *Nanofabrication: Principles, Capabilities and Limits*, Cham: Springer International Publishing, 2024, pp. 9–81. doi: 10.1007/978-3-031-62546-6_2.

[13] D. Jang, S.-G. Jung, S.-J. Min, and H.-Y. Yu, "Electrothermal Characterization and Optimization of Monolithic 3D Complementary FET (CFET)," *IEEE Access*, vol. 9, pp. 158116–158121, 2021, doi: 10.1109/ACCESS.2021.3130654.

[14] D. Das Sharma, G. Pasdast, S. Tiagaraj, and K. Aygün, "High-performance, power-efficient three-dimensional system-in-package designs with universal chiplet interconnect express," *Nat. Electron.*, vol. 7, no. 3, pp. 244–254, Feb. 2024, doi: 10.1038/s41928-024-01126-y.

[15] J. van Schoot *et al.*, "Next step in Moore's law: high NA EUV system overview and first imaging and overlay performance," *Journal of Micro/Nanopatterning, Materials, and Metrology*, vol. 24, no. 01, p. 011009, Dec. 2024, doi: 10.1117/1.JMM.24.1.011009.

[16] H. Arimura *et al.*, "Multi-Vt Gate Stack Technologies for Nanosheet and CFET Devices," in *2024 IEEE Silicon Nanoelectronics Workshop (SNW)*, IEEE, Jun. 2024, pp. 47–48. doi: 10.1109/SNW63608.2024.10639201.

[17] A. Bhardwaj *et al.*, "Comprehensive Analysis on Complementary FET," *IEEE Access*, vol. 13, pp. 82554–82572, 2025, doi: 10.1109/ACCESS.2025.3568134.

[18] K. Umstadter *et al.*, "EUV light source for high-NA and low-NA lithography," in *Optical and EUV Nanolithography XXXVI*, A. Lio and M. Burkhardt, Eds., SPIE, Apr. 2023, p. 43. doi: 10.1117/12.2657772.

[19] P. Onufryk and S. Choudhary, "UCIe: Standard for an Open Chiplet Ecosystem," *IEEE Micro*, vol. 45, no. 1, pp. 16–25, Jan. 2025, doi: 10.1109/MM.2024.3451532.

[20] D. Das Sharma, "PCI-Express: Evolution of a Ubiquitous Load-Store Interconnect Over Two Decades and the Path Forward for the Next Two Decades," *IEEE Circuits and Systems Magazine*, vol. 24, no. 2, pp. 47–61, 2024, doi: 10.1109/MCAS.2024.3373556.

[21] S. A. Chew, J. De Vos, and E. Beyne, "Wafer-to-wafer hybrid bonding at 400-nm interconnect pitch," *Nature Reviews Electrical Engineering*, vol. 1, no. 2, pp. 71–72, Feb. 2024, doi: 10.1038/s44287-024-00019-8.

[22] H. Peng, "Methodologies and Architectures for AI Inference Hardware: From Foundational Networks to Large Language Models," University of Washington, 2025.

[23] C. Y. Lee *et al.*, "3D Integrated Process and Hybrid Bonding of High Bandwidth Memory (HBM)," *Electronic Materials Letters*, vol. 21, no. 3, pp. 395–419, May 2025, doi: 10.1007/s13391-025-00557-9.

[24] J. A. Rosenau, A. W. Morales, S. S. Agili, and T. X. Tran, "Using Neural Networks for Far-End Crosstalk Compensation in High-Speed MIMO Channels," *IEEE Transactions on Signal and Power Integrity*, vol. 3, pp. 1–12, 2024, doi: 10.1109/TSIPI.2023.3335330.

[25] H. Kim *et al.*, "A Novel Interposer Channel Structure with Vertical Tabbed Vias to Reduce Far-End Crosstalk for Next-Generation High-Bandwidth Memory," *Micromachines (Basel).*, vol. 13, no. 7, p. 1070, Jul. 2022, doi: 10.3390/mi13071070.

[26] M. A. Al-Qadasi, L. Chrostowski, B. J. Shastri, and S. Shekhar, "Scaling up silicon photonic-based accelerators: Challenges and opportunities," *APL Photonics*, vol. 7, no. 2, p. 020902, Feb. 2022, doi: 10.1063/5.0070992.

[27] A. Di Meglio *et al.*, "Quantum Computing for High-Energy Physics: State of the Art and Challenges," *PRX Quantum*, vol. 5, no. 3, p. 037001, Aug. 2024, doi: 10.1103/PRXQuantum.5.037001.

[28] G. C. Hufbauer and M. Hogan, "Industrial policy through the CHIPS and Science Act: A preliminary report," Washington, DC, 2025.

[29] S. M. Shaji, L. Zhu, J. Yoon, and S. K. Lim, "A Comparative Study on Front-Side, Buried and Back-Side Power Rail Topologies in 3nm Technology Node," in *2023 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, IEEE, Aug. 2023, pp. 1–6. doi: 10.1109/ISLPED58423.2023.10244504.

[30] W. Rinehart and A. Kirchhoff, "The political economy of the CHIPS and Science Act," The Center for Growth and Opportunity. [Online]. Available: https://www.thecgo.org/research/the-political-economy-of-the-chips-and-science-act/

[31] B. Dachs, "The European Chips Act," 2023, *FIW - Research Centre International Economics, Vienna*.

[32] S. Ezell, "Assessing India's readiness to assume a greater role in global semiconductor value chains," 2024. [Online]. Available: https://itif.org/publications/2024/02/14/india-semiconductor-readiness/

[33] STMicroelectronics, "2023 Sustainability report," 2023. [Online]. Available: https://sustainabilityreports.st.com/sr23/

[34] Y. Zhou, Y. Li, and E. Ong, "Advancements in greenhouse gas emission reduction methodology for fluorinated compounds and N2O in the semiconductor industry via abatement systems," *Front. Energy Res.*, vol. 11, p. 1234486, Jan. 2024, doi: 10.3389/fenrg.2023.1234486.

[35] S. C. Song *et al.*, "System Design Technology Co-Optimization for 3D Integration at >5nm nodes," in *2021 IEEE International Electron Devices Meeting (IEDM)*, IEEE, Dec. 2021, pp. 22.3.1-22.3.4. doi: 10.1109/IEDM19574.2021.9720530.

[36] S. Sandhu, A. Zumeit, Z. Tian, V. Vinciguerra, and R. Dahiya, "Semiconductor manufacturing wastewater challenges and the potential solutions via printed electronics," *iScience*, vol. 28, no. 10, p. 113576, Oct. 2025, doi: 10.1016/j.isci.2025.113576.