STM JOURNALS

Journal of
Semiconductor Devices and Circuits

ISSN: 2455-3379
Volume 13, Issue 1, 2026
January–April
DOI (Journal): 10.37591/JoSDC

https://journals.stmjournals.com/josdc

Review JoSDC

Next-Generation Semiconductors: Making AI, 5G, and Beyond

V. Basil Hans*

Abstract

The current digital environment is being reshaped by next-generation semiconductors, which are powering revolutionary technologies, including advanced computing systems, 5G connectivity, and artificial intelligence. Next-generation semiconductors are at the heart of emerging technologies that are transforming the world today, such as artificial intelligence (AI), 5G communication, and advanced computing systems. As traditional silicon-based designs reach their limits in terms of size and performance, new materials, architectures, and manufacturing techniques are being developed to address the need for speed, efficiency, and scalability. New technologies, such as compound semiconductors, nanoscale transistors, and chiplet-based designs, are making it possible to process data more quickly, consume less power, and connect more devices. These developments are essential for data-intensive applications such as the Internet of Things (IoT), autonomous systems, and machine learning. Notwithstanding these advancements, issues including cost constraints, manufacturing complexity, and thermal management continue to pose major challenges. This article examines how semiconductor technology has evolved, highlights important advances that will improve next-generation performance, and discusses the challenges and opportunities that will shape the future of the semiconductor industry in an increasingly digital world.

Keywords: Artificial intelligence (AI), 5G technology, next-generation semiconductors, chiplet architecture, nanoscale transistors, compound semiconductors, and the internet of things (IoT)

INTRODUCTION

Semiconductors are the backbone of the modern economy and underpin many technological developments in areas such as telecommunications and artificial intelligence (AI). The semiconductor industry is working to meet the new demands that emerging technologies place on semiconductor materials [1, 2]. AI and 5G cellular technology—the next-generation of wireless standards—and their expected successors will create a need for specialized chips and systems that can process information more quickly and efficiently using advanced materials. Heterogeneous integration and explicit memory-computing systems can support a wide range of AI applications by linking specialized processors with high-bandwidth memory. For all phases of the AI processing pipeline and for 5G and beyond, high-speed data transmission requires high-frequency electrical and photonic components [3, 4].

Advanced semiconductor devices incorporate materials that allow signals to be processed faster and with less energy while also enabling new computing methods. Indium phosphide, gallium nitride, two-dimensional, semimetallic, and topological materials are some of the emerging materials that reduce scattering and energy loss. Quantum-dot, ferroelectric, phase-change, and vertical-hybrid-memory technologies improve on-chip memory operations. These technologies are necessary to satisfy the 5G requirements for ultra-reliable, low-latency communication, extensive machine-type communication, and enhanced mobile broadband [5, 6].

***Author for Correspondence**
V. Basil Hans
E-mail: vhans2011@gmail.com

Research Professor, Department of Commerce and Management and Humanities & Social Sciences, Srinivas University, Mangalore, Karnataka, India

Received Date: April 27, 2026
Accepted Date: April 30, 2026
Published Date: May 01, 2026

Citation: V. Basil Hans. Next-Generation Semiconductors: Making AI, 5G, and More Possible. Journal of Semiconductor Devices and Circuits. 2026; 13(1): 27–40p.

BASIC IDEAS ABOUT NEXT-GENERATION SEMICONDUCTORS

Next-generation semiconductor materials and device architectures enhance Moore's law by utilizing phenomena that transcend the limitations of the silicon bandgap. Two-dimensional atomic materials, such as graphene and transition metal dichalcogenides, are being investigated for low-power, flexible electronics and gain-based devices. Semiconductor devices for terahertz signal generation and detection are useful for radar and spectroscopic applications. III–V compounds and topological insulators are two classes of materials that show promise [7, 8]. Ceramic and solution-based metal–oxide semiconductors enable the fabrication of high-performance, large-area electronics for displays, sensors, and communications using roll-to-roll processes. Hybrid organic-inorganic perovskites exhibit rapid charge transport, tunable band gaps, and excellent photoluminescence efficiency. They can be used to manufacture displays with precisely defined pixels and advanced light-emitting devices.

The semiconductor industry is approaching the fundamental limits of silicon scaling. Future nodes are likely to face high costs and diminishing performance gains; hence, it is very important to identify new materials and device architectures. It is important to investigate new materials, device architectures, and the underlying physics. To replace silicon as the primary switching material, there must be a strong justification based on both architecture and economics [9, 10].

Semiconductor devices that can generate terahertz (THz) signals are important for radar and spectroscopic applications. Systems that can generate sub-terahertz signals are also important for modern telecommunications and terahertz spectroscopy. III–V compounds, such as indium phosphide and gallium nitride, and topological insulators, such as bismuth-antimony alloys, are emerging materials that could affect these applications.

Architecture and Materials for Devices

The most advanced computing applications have always pushed, challenged, and transformed technology and architecture. The future of industry and society as a whole now depends on the development of sophisticated semiconductors that enable AI, wideband telecommunications, and immersive multisensory media as these technologies evolve [11, 12]. Semiconductor solutions for these industries require specific features in the main aspects of semiconductor technology, such as materials and device architecture, capability scaling, fabrication and manufacturing methods, energy efficiency, and dependability. Therefore, materials and device architecture remain important in modern semiconductor technologies, while AI and telecommunications define their purpose and value.

Improvements in Fabrication and Manufacturing

Although fundamental physics sets limits on how small devices can become, breakthroughs in materials, device design, manufacturing, and packaging can keep semiconductor technology moving forward. These innovations make it possible to integrate many more transistors into a device, leading to significant gains in performance, power efficiency, and cost that meet the needs of the 5G and AI revolutions [13, 14]. Over the past few decades, the electronics industry has developed methods for fabricating individual transistors and complete integrated circuits at the nanometer scale. This has allowed electronic devices to become smaller and more capable.

Trends in Scaling and Theoretical Limits

The semiconductor industry has been driven by the macrotrend of increasing transistor counts, which has kept Moore's law relevant. Traditionally, the number of transistors has doubled approximately every two years. The latest 5 nm processors have approximately 50 billion transistors. As shown in Fig. 8, the scaling of feature sizes in commercial semiconductor technology and the associated constraints have created theoretical boundary conditions that help explain this exponential trend [15, 16]. The feature sizes for commercial silicon technology are now at the 5 nm node. However, new research is being conducted on 3D stacking methods, such as chiplets and through-silicon vias, and other approaches to heterogeneous integration. Currently, the concept of scaling can also refer to the number of chips or dies integrated into a single package [17, 18].

SEMICONDUCTOR SOLUTIONS FOR AI

AI, which encompasses everything from virtual assistants to self-driving cars, is paving the way for a new phase of computing and the next digital revolution. Semiconductor solutions are essential for accelerating the development of AI technologies and scaling AI algorithms to handle increasing training and inference workloads [19, 20]. For semiconductor solutions to support AI applications, they require computing accelerators capable of handling intensive computational workloads; memory-computing integration technologies that help overcome the memory wall; improved energy efficiency to address the power constraints of high-complexity accelerator models; and thermal management methods to overcome heat-sink limitations in systems or individual chips, enabling flexible integration of different types of chiplets [7, 8].

Accelerators and Computing with Different Types of Hardware

AI has enabled numerous advanced applications, including virtual assistants, visual perception, reinforcement learning, human–computer communication, text generation, and medical diagnostics. Different AI and deep learning approaches, models, and technologies have evolved rapidly. Researchers and innovators worldwide can use AI frameworks, models, and open-source code. Chip manufacturers, large digital platforms, and many businesses continue to invest heavily in AI research and development [21, 22]. To enable broader adoption of AI, machine learning systems need greater computing and storage capacity, as well as improved ease of deployment, flexibility, generalizability, and programmability. Therefore, AI and machine learning workloads require dedicated hardware accelerators.

Hardware accelerators, including graphics processing units (GPUs), application-specific integrated circuits (ASICs), field-programmable gate arrays (FPGAs), and video processing units (VPUs), have proven highly important in making AI and deep learning systems faster and more energy efficient. GPUs are the most common type of hardware used to train deep neural network models in data centers and cloud platforms. GPUs can process large amounts of data simultaneously because they can perform multiple tasks in parallel, but substantial effort is required to ensure that algorithms and model parameters perform optimally. Many real-time industrial applications use ASICs. ASICs enable ultrafast decision-making by providing low-latency inference [23, 24]. However, ASIC development is costly, time-consuming, and complex, with a return on investment that may be too low for many applications. FPGAs are a good choice because they balance flexibility, development time, performance, and cost. FPGAs can significantly enhance the performance and quality of various AI and deep learning applications, including model-loading optimization and mixed-precision computing. ASIC- and FPGA-based approaches, designs, and implementations often differ, making it difficult to share resources, skills, and expertise between architectures [25, 26].

Integration of Memory and Computing

Conventional approaches to semiconductor and AI system design have primarily focused on increasing voltage, power, and processing speed. However, as computational architectures evolve, greater attention must be given to how data are represented, stored, and accessed during processing, particularly for neural network training and inference tasks. Memory-compute integration is a computer architecture approach that brings memory and computational operations closer together to improve system performance and efficiency. It has demonstrated significant potential for accelerating AI workloads.

Managing Heat and Energy Use

To achieve the best possible performance with minimum energy consumption, several design considerations must be addressed. These include selecting low-power circuits and developing architectures that minimize energy consumption while sustaining the required workloads. AI, machine learning (ML), and deep learning workloads are increasingly deployed in specialized computing domains rather than in conventional general-purpose computing environments, particularly in ultra-low-power and microcontroller-based applications. Integrated circuits (ICs) developed for smart sensor

systems generally follow these principles. However, ICs, circuit boards, enclosures, and operating environments introduce platform-specific constraints that differ from those of conventional computing systems. Because compact smart devices are subject to stringent size and area limitations, designing systems that deliver adequate performance while minimizing energy consumption remains a major challenge.

Although the human brain can perform tasks such as image classification and anomaly detection using approximately 2–20 W of power, deep learning workloads can require substantially more energy. Consequently, energy-efficient AI applications designed for resource-constrained and compact devices are increasingly important. These applications include deep-audio processing, text processing, time-series trend detection, and multisensor data fusion, together with early-threshold mechanisms for true edge intelligence. Such applications require energy-efficient chip architectures for smart sensors that can operate across diverse platforms and application stacks, ranging from academic research systems to the Internet of Things (IoT). The industry therefore requires energy-efficient computing architecture rather than continued reliance on increasingly large, infrastructure-centric platforms. Distributed, highly scalable, and architecturally heterogeneous deep learning systems can be implemented across the system-on-chip (SoC) spectrum, ranging from GPUs and specialized low-power hardware/software architectures to advanced material-based SoCs and microcontrollers (MCUs), as well as application-specific AI-enabled edge platforms. These approaches can reduce energy consumption across diverse applications and support the emerging transition toward intelligent audio and other forms of edge sensing.

Computing systems have evolved from primarily CPU-based architectures to heterogeneous platforms comprising CPUs, GPUs, dedicated accelerators, and FPGAs. Edge devices such as Raspberry Pi and Google's Coral Tensor Processing Unit (TPU) now provide heterogeneous computing capabilities. Smaller devices must also address unavoidable thermal and thermodynamic constraints, particularly at advanced technology nodes. Designing energy-efficient systems is challenging because of domain-specific processing requirements, architectural heterogeneity, system-level constraints, and highly constrained form factors suitable for TinyML applications. Architecting, designing, implementing, and verifying such systems require multidisciplinary expertise spanning hardware design, process technology, microarchitecture, software, firmware, operating systems, thermal management, and workload characterization. Researchers have investigated energy-saving techniques across hardware, circuit, microarchitectural, and process technology levels. They have also modeled and simulated power consumption, performance, and thermal behavior at different levels of hardware and system design.

5G AND BEYOND SEMICONDUCTOR SOLUTIONS

Next-generation semiconductor devices based on advanced materials and novel device architectures are essential for enabling the wireless revolution by significantly improving the performance of radio-frequency front-end components and the computational capabilities required for AI-assisted 5G systems. AI-enabled 5G networks require substantial resources for edge processing, data storage, caching, and routing, in addition to supporting very high data rates and low latency. Wireless systems must move beyond conventional microwave technologies and traditional device-scaling approaches toward new architectures that exploit terahertz-frequency transmission. This transition is particularly important because the available lower-frequency spectrum is becoming increasingly congested as the demand for wireless capacity continues to grow. Advanced backhaul technologies can further improve signal quality and reduce power consumption. Addressing emerging technology stacks and the associated challenges in thermal management, signal integrity, and performance-to-form-factor trade-offs require advanced heterogeneous and hybrid packaging technologies that integrate multiple chips into a single module. Such approaches are important for enabling architectures supporting ultra-massive multiple-input multiple-output (MIMO) systems, quantum communication, and reconfigurable intelligent surfaces (RISs).

Front-End Components with High Frequencies

Current antennas for 5G and earlier wireless technologies are employed in both massive MIMO and conventional MIMO configurations. These architectures directly connect radios and antennas, which limits the number of radios and increases the length of the silicon signal chain. Photonic ICs (PICs) are being used to generate radio-frequency (RF) signals in SoC solutions for 6G and beyond. These chips use optical frequency combs and photodetection to generate microwave-to-terahertz frequencies that are much higher than those achievable with conventional electrical techniques. SoC solutions that integrate silicon photonics with semiconductor platforms, such as silicon–germanium Bipolar Complementary Metal-Oxide-Semiconductor (SiGe BiCMOS), make RF systems more compact.

To meet the diverse requirements of 6G applications, concepts such as cognitive radio use AI to optimize waveforms and improve performance through resource optimization, fault prediction, and intelligent scheduling [15]. These methods maximize the efficiency of dynamic spectrum-sharing networks while complying with national regulatory requirements. Graph neural networks (GNNs) add cognitive capabilities to RF chains, enabling open architectures and continual learning. This is because GNNs can adapt to topological changes in network graphs without requiring the redesign of application-specific algorithms. Advanced AI-based waveform optimization exploits the interdependence of spectrum, power, and time resources to improve communication efficiency across heterogeneous wireless networks..

Backhaul and Edge Processing

The rapid deployment of 5G technology is increasing the demand for high-speed data center interconnection (DCI) and mobile edge computing (MEC) solutions. To address the growing requirements for high-speed backhaul data transmission and low-latency processing at the network edge [16], DCI and MEC semiconductor technologies must achieve greater efficiency. 5G New Radio (NR) introduces access to higher-frequency spectrum, including the millimeter-wave (mmWave) bands, generally spanning 30–300 GHz. These bands are intended to provide users with data rates of several gigabits per second. Fiber-optic links are typically used to transport wireless mmWave signals over distances exceeding a few hundred meters. Therefore, efficient DCI solutions are required to maintain reliable connectivity between the radio access network (RAN) and the core network.

5G supports low-latency applications that require data processing to be performed close to the user equipment (UE). MEC solutions help reduce latency by moving computational resources closer to end users. Coarse wavelength-division multiplexing (CWDM) is currently one of the commonly used optical technologies for backhaul and DCI applications. It typically relies on optical–electrical–optical (OEO) conversion for data transmission. Although OEO-based approaches are effective for conventional DCI applications, they can impose limitations on transmission distance and network scalability. To address these limitations, high-order modulation formats using analog optical signals (AOSs) for extended-distance DCI applications and AOS-based analog baseband radio-frequency circuits (BBRFs) for MEC applications have been proposed and demonstrated. These approaches can improve data capacity and spectral efficiency while providing greater flexibility in network deployment.

Technologies for Packaging and Connecting

The 5G infrastructure is required to support an increasingly diverse range of workloads driven by the growing demand for always-on connectivity, high-definition video, and advanced AI capabilities. These workloads include data processing, storage, and transmission across fronthaul, backhaul, and edge networks [17]. Silicon-based electronic systems must operate efficiently in these environments while simultaneously supporting diverse communication, sensing, and processing functions. As interconnect latency becomes increasingly important, new packaging and interconnect technologies are being developed to enable the integration of heterogeneous silicon dies. Unlike traditional systems that rely on pre-tested dies or chips connected through conventional interconnects, optimizing the design and placement of nearest-neighbor interconnects and performing electrical and thermal routing at the package level can significantly improve overall system performance.

TRUST, SECURITY, AND RELIABILITY IN ADVANCED SEMICONDUCTORS

Increasingly, the performance, dependability, and security of digital electronic systems depend not only on their constituent materials but also on their components, applications, architectures, and interconnections. Nevertheless, modern semiconductor technologies and their associated design methodologies continue to face significant challenges at both the material and device levels. As SoCs become increasingly complex, incorporating heterogeneous architectures and numerous interconnections, data integrity, privacy, and provenance are becoming important design objectives. The geographically distributed nature of global semiconductor supply chains further increases their vulnerability and can affect the economic and commercial viability of AI and fifth-generation (5G) technologies.

Exposure to ionizing radiation, such as that encountered in space systems or during nuclear and radiological events, can seriously impair the performance, reliability, and functionality of microelectronic circuits and systems. Ionizing radiation can induce single-event transients (SETs) in sensitive circuit nodes, particularly in devices fabricated using advanced semiconductor technologies. Energetic particle interactions can also cause displacement damage within the crystal lattice. Designing ICs with fault-tolerant architectures and radiation-hardening techniques can substantially improve their resistance to these external stressors.

The increasing sophistication of hardware Trojans, logic manipulation, and other invasive attacks targeting modern SoCs across diverse application sectors represents a significant global security concern. Attacks on commercially valuable designs can compromise data privacy, expose intellectual property (IP), and undermine customer trust in technology vendors. IP escrow mechanisms, which are primarily intended to address risks associated with the loss, transfer, or availability of IP, cannot mitigate all of these threats. Therefore, robust hardware-security primitives, such as physical unclonable functions (PUFs), secure authentication mechanisms, and secure key-storage systems, have become increasingly important for SoCs that interface with external memory. Such mechanisms help protect not only the security architecture itself but also, critically, the underlying design data and IP that constitute a core component of the technology value chain.

The emergence of the digital-twin paradigm—the digital representation of a physical asset—offers significant potential for accelerating the adoption of advanced semiconductor technologies while improving the monitoring, accessibility, traceability, and portability of the design data associated with these technologies. Comprehensive digital twins that extend throughout the asset lifecycle, including the monitoring and management of deployed systems, and that incorporate advanced AI systems for analyzing components, devices, and operating scenarios at global scale could further stimulate advances in AI-enhanced semiconductor components and systems. These concepts are currently being investigated by various government, industry, and research organizations.

Hardening Against Radiation and Making Mistakes

Advanced semiconductor devices are increasingly used in space, medical, aerospace, and industrial applications, where they must operate under harsh environmental conditions. These conditions may include high temperatures, mechanical stress, humidity, and ionizing radiation. This section discusses radiation-induced voltage transients and other radiation effects in advanced semiconductor technology nodes. It also examines radiation-hardened large-scale integrated (LSI) circuits designed to improve resistance to radiation-induced failures, particularly single-event upsets (SEUs). Radiation sources, including protons, neutrons, alpha particles, beta particles, and gamma rays, can adversely affect semiconductor devices, resulting in increased off-state leakage current, reduced gain, and degradation of overall device performance.

Radiation effects in semiconductor devices can be broadly classified into single-event effects (SEEs), including (1) single-event latch-up (SEL), (2) SEU, and (3) single-event transient (SET), as well as total ionizing dose (TID) effects. These mechanisms differ in their physical origins, electrical signatures, and temporal characteristics. SEL can induce a persistent high-current state and may require power cycling

to recover, whereas an SEU causes a temporary or permanent change in the logical state of a memory or digital circuit. An SET is characterized by a transient voltage or current pulse generated by the passage of a high-energy particle through a sensitive region of a device. In advanced LSI circuits, appropriate circuit-level and layout-level protection can significantly reduce the susceptibility to SEUs, SEL, and SETs. Consequently, SET characterization is an important approach for evaluating the radiation sensitivity of advanced technology nodes.

Researchers have investigated the radiation hardness and radiation-tolerant design of LSI circuits exposed to heavy-ion radiation. Simulation and experimental studies involving more than 20,000 devices fabricated using 28-nm complementary metal–oxide–semiconductor (CMOS) technology have demonstrated the effectiveness of radiation-tolerant designs. Radiation-hardened clock-tree structures, combinational logic, and analog-cell architectures can improve circuit resistance to radiation-induced disturbances. Such radiation-hardened LSI circuits are particularly suitable for space applications, where exposure to energetic particles can significantly affect electronic-system reliability. The development of technology libraries and radiation-tolerant circuit architectures can further facilitate the implementation of these capabilities in advanced semiconductor technologies.

Researchers have also investigated heavy-ion-induced SETs in advanced ICs. Devices susceptible to heavy-ion-induced soft errors have been examined in the pulse-generation paths of double-sampling analog-to-digital converters (ADCs). By sharing power resources with selected circuit components that are relatively insensitive to hold-time-violation pulses, the effects of these transients can be reduced without substantially increasing power consumption or circuit area. In a clocked comparator employing a double-sampling topology, the observed frequency of SETs was reduced by approximately two orders of magnitude. These findings demonstrate that circuit-level architectural techniques can substantially improve radiation tolerance while maintaining the area and power constraints of advanced semiconductor systems.

Foundations of Hardware Security

The rapid development of AI, fifth-generation (5G) connectivity, and the IoT has placed increasing demands on semiconductor technologies. AI accelerators require extremely high computational performance, memory bandwidth, and energy efficiency to process increasingly large datasets and models with growing numbers of parameters. 5G applications require high-frequency communication circuits, edge and near-device computing solutions, and high-speed interconnects capable of supporting data rates exceeding 100 Gb/s. Meeting these requirements necessitates the development of diverse semiconductor technologies and device architectures. These technologies include advanced semiconductor materials, two-dimensional (2D) materials, novel three-dimensional (3D) integration schemes, emerging device architectures, and heterogeneous integration at both the chip and chiplet levels. Next-generation semiconductor technologies are enabling emerging applications in AI, 5G, and IoT while extending the principles of Moore's law beyond conventional transistor scaling and feature-size reduction. This article provides an overview of semiconductor technologies and solutions developed to address the performance, energy-efficiency, bandwidth, and integration requirements of these emerging applications.

Resilience in the Lifecycle and Supply Chain

Semiconductors are becoming increasingly important for AI and 5G infrastructure. However, emerging applications and evolving market requirements indicate that performance, economic, and architectural demands are changing rapidly. Energy-efficient computing remains a critical requirement, particularly for accelerating large deep learning models. Increasingly, heterogeneous computing platforms are integrating multiple types of accelerators to optimize computational workloads, power consumption, and overall performance. The number and complexity of AI inference models continue to increase, creating significant challenges associated with memory access and data movement. Bringing memory closer to the processing units is an effective approach to mitigating these challenges. Power distribution and thermal management also remain important considerations. Efficient clock-

management techniques and localized cooling solutions can help maintain an appropriate balance between computational performance and thermal constraints.

Next-generation wireless applications are being enabled by advances in semiconductor technologies. 5G and beyond-5G systems require operation at increasingly higher frequencies, making front-end components such as low-noise amplifiers (LNAs), power amplifiers (PAs), mixers, and phase-locked loops (PLLs) increasingly important. As the physical separation between antennas, RF front ends, and processing units increases, efficient backhaul, fronthaul, and edge computing solutions become essential. Specific applications may also require line-of-sight communication links and high-resolution sensor fusion. Integrated photonics and high-speed RF interconnects enable high-capacity data transmission for bandwidth-intensive 5G applications. The continued advancement of 5G is also being supported by innovative packaging and connectivity technologies, including chiplet-based architectures, which enable higher levels of integration, scalability, and system functionality..

THINGS TO THINK ABOUT: ECONOMICS, THE ENVIRONMENT, AND POLICY

The economics of semiconductor development indicate that costs are becoming increasingly differentiated across technology nodes and platforms, while technological leadership can provide companies with a significant competitive advantage. For example, transitioning from a mature technology node to a subsequent-generation node may result in an approximately 50% increase in average revenue, accompanied by only a 40% increase in incremental costs. This revenue advantage can be considerably greater for next-generation materials such as gallium nitride (GaN). Compared with silicon-based technologies, GaN can potentially generate more than four times the average revenue increase in certain applications because of its novel device architectures, topologies, and application opportunities. Consequently, emerging markets are increasingly focusing on technologies that extend beyond conventional silicon scaling to achieve new sources of economic value. This shift is driven by the expectation that emerging technologies can achieve competitive cost structures while providing capabilities that conventional silicon-based approaches cannot readily deliver.

Another important consideration in the future development of semiconductor technologies is their environmental impact. Improving the energy and material efficiency of semiconductor devices is becoming increasingly important as the energy requirements of data centers continue to grow and the semiconductor supply chain faces increasing resource and resilience challenges. Semiconductor manufacturers therefore have strong incentives to accelerate time-to-market not only for new chips but also for emerging technological concepts. These requirements are driving the development of new technology platforms encompassing advanced materials, fabrication processes, device architectures, packaging techniques, and manufacturing methodologies.

For semiconductor strategies to be successful, they must be supported by a strong foundation of shared knowledge, international collaboration, and effective cooperation among industry, academia, and government stakeholders. Such strategies should also consider and mitigate unintended geopolitical consequences that could create unequal technological advantages or increase dependencies among countries. Long-term sustainability in semiconductor materials and manufacturing processes requires active international cooperation in advanced technologies, including photonics, heterogeneous integration, and innovative packaging. Developing countries continue to face limited access to advanced equipment and facilities for materials and nanotechnology research. Nevertheless, next-generation materials and device architectures offer significant opportunities through the development of specialized materials, application-specific architectures, and emerging semiconductor technologies.

Cost of Ownership and Market Trends

The cost of adopting and maintaining advanced semiconductor technologies is a key factor influencing the current market dynamics and the future development of the semiconductor industry. The

total cost of ownership (TCO) encompasses not only the capital expenditure required to introduce new technologies but also net cash outflows, anticipated revenues, and operational-efficiency measures throughout the active lifecycle of a device or technology. Making informed strategic decisions regarding technology adoption therefore requires a comprehensive understanding of market acceptance, technology roadmaps, investment requirements, and commercialization timelines. By these measures, silicon-based technologies maintained a competitive advantage over emerging semiconductor technologies during the 1990s, when the net cash outflow associated with adopting alternative technologies increased rapidly. By the early 2020s, however, time-to-volume and time-to-revenue had become increasingly important indicators of technological competitiveness, alongside conventional cost considerations. These factors reduced the immediate economic pressure associated with transitioning toward post-silicon technologies and contributed to the continued dominance of silicon-based technologies over the subsequent two decades.

Sustainability and the Material Footprint

Heterogeneous integration (HI) is an advanced approach to integrated-circuit design that combines different types of devices, such as logic, memory, and sensors, within a single package or system. This approach differs from conventional monolithic integration, in which an entire chip is typically fabricated using a single semiconductor technology or process. The semiconductor industry has a substantial carbon footprint (CFP) arising from the entire computing ecosystem, ranging from semiconductor manufacturing to data centers. The information and communication technology (ICT) sector already contributes significantly to global carbon emissions and is projected to become an increasingly important source of emissions. In addition, the embodied carbon footprint associated with semiconductor design and fabrication can constitute a substantial proportion of the overall carbon footprint of low-power computing devices. Therefore, achieving sustainable and environmentally responsible computing requires consideration of both embodied and operational carbon footprints.

HI enables the development of high-density, high-bandwidth systems based on chiplets. These architectures can provide favorable cost–power–performance trade-offs, making them particularly attractive for sustainable very-large-scale integration (VLSI) systems. Chiplet-based HI can reduce individual die sizes, potentially improving manufacturing yield and reducing silicon waste. It also enables the reuse of pre-designed and validated chiplets across multiple products, thereby reducing design effort and potentially lowering the environmental impact associated with semiconductor development and manufacturing.

However, new modeling and evaluation methodologies are required to accurately estimate the carbon footprint of emerging multi-die, heterogeneous, chiplet-based systems. To address this need, ECO-CHIP has been proposed as a tool for estimating the embodied and operational carbon footprints, as well as the total carbon footprint, of advanced heterogeneous architectures within evolving chiplet-based VLSI technologies. Comprehensive evaluations indicate that appropriately designed heterogeneous architectures have considerable potential to reduce carbon emissions compared with conventional monolithic dies. As an open-source tool, ECO-CHIP can facilitate broader research and adoption of carbon-aware semiconductor design methodologies and contribute to the development of more sustainable and environmentally responsible VLSI systems.

International Cooperation, Standards, and Rules

The convergence of communication technologies and AI, together with the need to address global challenges such as climate change, pandemics, and energy shortages, highlights the need for significant social and economic transformation. These challenges provide strong motivation for greater collaboration in the development of standards, technologies, and cooperative frameworks. Industries rely on standards and regulatory frameworks to ensure interoperability, facilitate the global deployment of technologies, and prevent fragmentation across markets and technological ecosystems. In this

context, sixth-generation (6G) connectivity is expected to facilitate more seamless interactions among people, machines, and intelligent systems, potentially transforming the ways in which individuals, organizations, and societies communicate and interact [23].

NEW USES AND WHAT THE FUTURE HOLDS

Next-generation accelerators, memory technologies, and RF front-end components are poised to transform edge AI, autonomous vehicles, and communication systems beyond fifth-generation (5G) networks [1]. Most AI applications are currently based on algorithms deployed on computing infrastructure in data centers and cloud platforms. However, the growing demand for real-time processing, low latency, privacy, and reduced communication overhead is driving the development of energy-efficient AI and ML solutions for edge devices, including smartphones, IoT devices, and autonomous vehicles. Advanced semiconductor technologies are expected to play a central role in enabling this transition.

The semiconductor industry is actively developing quantum technologies, neuromorphic computing, edge AI, and other emerging computing paradigms to support the continued evolution of AI. First, the practical realization of quantum advantage remains challenging because current qubit-based systems are highly susceptible to errors, while limitations in materials, device fabrication, scalability, and quantum algorithms continue to constrain their development. Nevertheless, an increasing number of universities and research institutions are investigating semiconductor-based quantum devices and components, demonstrating continued progress toward practical quantum computing systems. Second, neuromorphic technologies, which exploit physical mechanisms and hardware architectures inspired by the organization and operation of biological neural systems, are receiving considerable attention. Their ability to process temporal information and exploit time-dependent correlations offers potential advantages for perception and edge intelligence applications. These developments also motivate fundamental research into circuit architectures capable of efficiently implementing ML and deep learning (DL) operations, thereby exposing the limitations of conventional computing architectures and semiconductor technologies.

Beyond edge AI, integrated infrastructures that combine communication, sensing, computing, and actuation are emerging as important components of sixth-generation (6G) and beyond-6G systems. Both academia and industry are actively investigating a broad range of enabling technologies, creating significant opportunities for semiconductor innovation and commercialization. Emerging 6G concepts extend beyond the capabilities and architectural assumptions of current 5G systems and may require advances in high-frequency devices, integrated photonics, HI, advanced packaging, AI-enabled radio systems, sensing technologies, and energy-efficient edge computing.

AI at Edge and Autonomous Systems

For AI to function effectively in edge and autonomous systems, devices must support high-speed, low-power edge computing. In applications such as the IoT and vehicular networks, where data are generated at the edge, local processing is essential. Dependence on remote cloud servers, data transfer between edge nodes, and transmission delays can make cloud-based processing unsuitable for time-sensitive applications. To address privacy concerns associated with cloud-based data sharing and limitations in local infrastructure, robust edge and on-device AI systems can process, analyze, and extract meaningful information from data securely and efficiently. Deep neural networks (DNNs) are widely used for feature extraction and decision-making in AI applications. However, deploying neural networks in edge and autonomous systems presents significant challenges because of limitations in computational resources, energy availability, memory capacity, and communication bandwidth. Therefore, it is essential to develop devices capable of performing AI-enabled real-time tasks at the edge in a more efficient, energy-conscious, and adaptable manner.

AI is expected to increasingly shift from cloud-based platforms toward edge environments over the coming years, particularly in applications such as mobile Internet, autonomous vehicles, smart

manufacturing, and industrial AI. This trend further emphasizes the need to develop dedicated edge-AI chips capable of supporting intelligent analysis and real-time decision-making. The open-source community can provide researchers and developers with access to a wide range of DL frameworks, models, tools, and techniques, facilitating the development and customization of AI models for edge devices. This can make AI-edge systems more adaptable to diverse application requirements. However, the requirements of AI applications in edge and autonomous systems differ substantially from those of general-purpose AI applications. Consequently, conventional AI accelerators and chips designed for data center or cloud environments may not be suitable for edge applications, highlighting the need for application-specific and energy-efficient edge-AI hardware.

Computing that is Ready for Quantum and Neuromorphic Computers

Computational speed and energy efficiency are becoming increasingly important as billions of connected IoT devices and 5G-enabled smart systems generate and transmit vast amounts of data. The growing demand for faster and more energy-efficient computing, together with the need to process information closer to its source, has motivated the development of alternative computing paradigms. Among these, neuromorphic computing has attracted considerable attention. Inspired by the structure and information-processing mechanisms of the human brain, neuromorphic computing aims to achieve high computational efficiency and low energy consumption compared with conventional computing architectures.

A related approach seeks to exploit the unique computational capabilities of quantum systems. Quantum-ready computing systems, which are designed to interface with and exploit quantum resources, have attracted significant interest because of their potential to enable new computational capabilities and accelerate selected workloads. For example, quantum neural networks (QNNs) can employ quantum-enhanced kernels for certain ML tasks, potentially improving the efficiency of training and inference. Quantum time-multiplexed echo-state networks can also be implemented using different quantum platforms, including systems based on qubits, bosonic states, and many-body quantum systems, providing a flexible framework for reservoir computing.

Photonic computing and emerging materials and devices based on topological phenomena represent two promising research directions for advancing neuromorphic and quantum-ready computing technologies [24–26]. These approaches offer potential advantages in terms of parallel processing, energy efficiency, high-speed information transmission, and the integration of unconventional computational mechanisms.

Beyond 5G Infrastructure and 6G Horizons

Based on the experience gained from previous 5G deployments, the period leading up to 2030 and beyond is expected to witness extensive co-design and exploration of transformative technologies. The rapid emergence and adoption of AI capabilities across a wide range of practical applications in recent years have considerably expanded the design space for future wireless systems. Future wireless mobile communication networks are expected to provide a platform for developing and testing a new generation of technologies, commonly referred to as sixth-generation (6G) systems. Several concepts have already been proposed for next-generation mobile communication systems, including new services, network architectures, key performance indicators (KPIs), and the use of AI to support or integrate intelligent functionalities into the communication system [13,14].

The development of 6G mobile communication systems requires not only entirely new technological concepts but also a reconsideration of the design constraints and assumptions underlying current mobile communication systems, including those associated with the global deployment of 5G. Relaxing these constraints can expand the co-design space and facilitate the exploration of more innovative solutions. Research into beyond-5G (B5G) and 6G technologies continues to build on the fundamental service frameworks, user requirements, usage scenarios, performance constraints, and wireless-access capabilities established by current mobile communication systems. This evolutionary approach

provides a foundation for developing more flexible, intelligent, and integrated communication networks capable of supporting emerging applications and services.

CONCLUSION

The semiconductor industry is entering a new decade of innovation driven by the rapidly growing demand for AI, fifth-generation (5G) telecommunications, and other emerging applications. To sustain and expand the benefits of microelectronics, continued investment in advanced semiconductor technologies, system architectures, and integration approaches is essential. Several factors are expected to make the coming decade a critical period for semiconductor innovation.

AI is one of the primary drivers of this transformation. The rapid development of AI, large language models (LLMs), generative AI, and multimodal applications has significantly changed the requirements for computing performance, memory capacity, connectivity, and energy efficiency compared with those of mainstream workstation applications a decade ago. New multimodal LLMs and other advanced models with increasingly sophisticated capabilities are emerging rapidly. The unprecedented scale and complexity of these software systems are creating new architectural and process technology requirements for semiconductor devices. These developments provide opportunities to push technological boundaries, optimize trade-offs across chip, system, and application lifecycles, explore advanced materials, and investigate medium- and long-term alternatives to conventional silicon and III–V semiconductor technologies [2].

Another major driver is the widespread deployment of 5G, cloud computing, and other connected technologies. The continued expansion of 5G networks is influencing individual, industrial, and societal applications more extensively than initially anticipated. Intelligent and self-learning agents are finding applications across diverse sectors, including manufacturing, hospitality, healthcare, e-commerce, and maintenance services, while research continues toward increasingly capable autonomous agents.

At the physical-layer level, advances in channel equalization, high-frequency electronics, and gigahertz- and terahertz-range operation are enabling flexible, high-order modulation techniques in increasingly congested frequency bands. The evolution of smart-grid technologies and the extension of time-sensitive networking (TSN) are also important for supporting ultra-reliable low-latency communication (URLLC) under increasingly stringent performance requirements. At the same time, fourth-generation (4G) Long-Term Evolution (LTE) technologies continue to provide important connectivity infrastructure alongside 5G.

Emerging semiconductor architectures are also targeting substantial improvements in computational efficiency. These include accelerator–processor architectures, advanced memory and interconnect technologies, and HI approaches designed to reduce energy consumption and improve computational throughput. Such developments are particularly important for next-generation AI workloads, where achieving high performance within stringent power and thermal constraints remains a major challenge.

REFERENCES

1. Chataut R, Nankya M, Akl R. 6G networks and the AI revolution-exploring technologies, applications, and emerging challenges. Sensors. 2024;24(6):1888. doi:10.3390/s24061888. PMID:38544151.
2. Gottscho RA, Levine EV, Liu TJ, McIntyre PC, Mitra S, Murmann B, et al. Innovating at speed and at scale: A next generation infrastructure for accelerating semiconductor technologies. arXiv [Preprint]. 2022 Mar 7. arXiv:2204.02216.
3. Migliato Marega G, Wang Z, Paliy M, Giusi G, Strangio S, Castiglione F, et al. Low-power artificial neural network perceptron based on monolayer MoS2. ACS Nano. 2022;16(3):3684-3694. doi:10.1021/acsnano.1c07065. PMID:35167265.
4. Feng C, Wu W, Liu H, Wang J, Wan H, Ma G, et al. Emerging opportunities for 2D materials in neuromorphic computing. Nanomaterials. 2023;13(19):2720. doi:10.3390/nano13192720. PMID:37836361.

5. Bathe M, Chrisey LA, Herr DJC, Lin Q, Rasic D, Woolley AT, et al. Roadmap on biological pathways for electronic nanofabrication and materials. Nano Futures. 2019;3(1):012001. doi:10.1088/2399-1984/aaf7d5.
6. Sun Y, Agostini NB, Dong S, Kaeli D. Summarizing CPU and GPU design trends with product data. arXiv [Preprint]. 2019 Nov 26. arXiv:1911.11313.
7. James AP. The why, what, and how of artificial general intelligence chip development. IEEE Trans Cogn Dev Syst. 2022;14(2):333-347. doi:10.1109/TCDS.2021.3069871.
8. Zenbaa N, Abert C, Majcen F, Kerber M, Serha RO, Knauer S, et al. Experimental realisation of a universal inverse-design magnonic device. arXiv [Preprint]. 2024 Mar 26. arXiv:2403.17724.
9. Reuther A, Michaleas P, Jones M, Gadepally V, Samsi S, Kepner J. Survey of machine learning accelerators. 2020 IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA. 2020; pp. 1-12. doi:10.1109/HPEC43674.2020.9286149.
10. Yang Z, Ji S, Chen X, Zhuang J, Zhang W, Jani D, et al. Challenges and opportunities to enable large-scale computing via heterogeneous chiplets. 2024 29th Asia and South Pacific Design Automation Conference (ASP-DAC), Incheon, Republic of Korea. 2024; pp. 765-770. doi:10.1109/ASP-DAC58780.2024.10473961.
11. Haensch W, Raghunathan A, Roy K, Chakrabarti B, Phatak CM, Wang C, et al. A co-design view of compute in-memory with non-volatile elements for neural networks. arXiv [Preprint]. 2022 Jun 3. arXiv:2206.08735.
12. Muralidhar R, Borovica-Gajic R, Buyya R. Energy efficient computing systems: Architectures, abstractions and modeling to techniques and standards. ACM Comput Surv. 2022;54(11s):1-37. doi:10.1145/3511094.
13. Letaief KB, Chen W, Shi Y, Zhang J, Zhang YJA. The roadmap to 6G: AI empowered wireless networks. IEEE Commun Mag. 2019;57(8):84-90. doi:10.1109/MCOM.2019.1900271.
14. Xue R, Tan J, Shi Y. Exploration and application of AI in 6G field. arXiv [Preprint]. 2022 Jul 27. arXiv:2207.13382.
15. Cousik T, Shafin R, Zhou Z, Kleine K, Reed J, Liu L. CogRF: A new frontier for machine learning and artificial intelligence for 6G RF systems. arXiv [Preprint]. 2019 Sep 15. arXiv:1909.06862.
16. Peltonen E, Bennis M, Capobianco M, Debbah M, Ding A, Gil-Castiñeira F, et al. 6G white paper on edge intelligence. arXiv [Preprint]. 2020 Apr 30. arXiv:2004.14850.
17. Noor R, Kottur HR, Craig PJ, Biswas LK, Khan MS, Varshney N, et al. US microelectronics packaging ecosystem: Challenges and opportunities. arXiv [Preprint]. 2023 Oct 18. arXiv:2310.11651.
18. Schoof G, Kraemer R, Jagdhold U, Wolf C. Fault-tolerant design for applications exposed to radiation. In: Ouwehand L, editor. Proceedings of DASIA 2007: Data Systems in Aerospace Conference; 2007 May 29-Jun 1; Naples, Italy. ESA SP-638. Noordwijk (Netherlands): European Space Agency; 2007.
19. Venkatesha S, Parthasarathi R. A survey of fault models and fault tolerance methods for 2D bus-based multi-core systems and TSV based 3D NOC many-core systems. arXiv [Preprint]. 2022 Mar 15. arXiv:2203.07830.
20. Mullen E, Morris MA. Green nanofabrication opportunities in the semiconductor industry: A life cycle perspective. Nanomaterials. 2021;11(5):1085. doi:10.3390/nano11051085. PMID:33922231.
21. Hsiao HC, Wen HC, Nakano M. A nonlinear growth analysis of integrated device manufacturers' evolution to the nanotechnology manufacturing outsourcing. Int J Eng Technol Innov. 2012;2(2):150-162.
22. Lin X, Kundu L, Dick C, Velayutham S. Embracing AI in 5G-advanced toward 6G: A joint 3GPP and O-RAN perspective. IEEE Commun Stand Mag. 2023;7(4):76-83. doi:10.1109/MCOMSTD.0005.2200070.
23. Wang D. Pushing the boundaries: An interview with Dae-Hyun Kim on terahertz devices. Natl Sci Rev. 2024;11(3):nwae013. doi:10.1093/nsr/nwae013. PMID:38327852.

24. Finocchio G, Incorvia JAC, Friedman JS, Yang Q, Giordano A, Grollier J, et al. Roadmap for unconventional computing with nanotechnology. Nano Futures. 2024;8(1):012001. doi:10.1088/2399-1984/ad299a.
25. Li R, Gong Y, Huang H, Zhou Y, Mao S, Wei Z, et al. Photonics for neuromorphic computing: Fundamentals, devices, and opportunities. Adv Mater. 2025;37(2):e2312825. doi:10.1002/adma.202312825. PMID:39011981.
26. Assi DS, Huang H, Karthikeyan V, Theja VCS, de Souza MM, Xi N, et al. Quantum topological neuristors for advanced neuromorphic intelligent systems. Adv Sci (Weinh). 2023;10(24):e2300791. doi:10.1002/advs.202300791. PMID:37340871.

