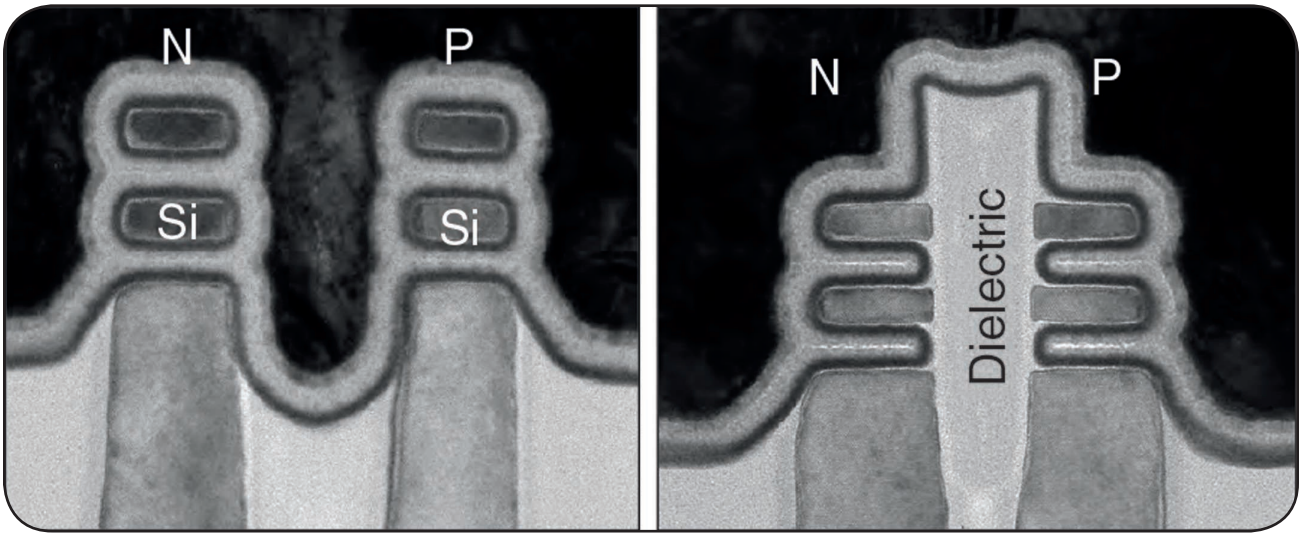
Nadine Collaert

Advancements in IC Technologies

A look toward the future



SCALING ROADMAP LOGIC TRANSISTOR

The invention of the transistor 75 years ago paved the way for a technological revolution that has transformed our world in ways that were once unimaginable. From the PC to the Internet and from mobile phones to artificial intelligence (AI), transistors have powered the exponential growth of the digital age. Yet, even as we celebrate this milestone in semiconductor history, we find ourselves at a critical inflection point.

Introduction

For more than a decade now, the semiconductor industry has been grappling with the limitations of CMOS technology and the implications of Moore's law reaching its limits. The dimensional scaling that once propelled progress in transistor technology is no longer a reliable predictor of future advancements, and we are facing severe power and speed roadblocks in our complex systems. As a result, it is becoming increasingly clear that we cannot rely on a single technology to fulfill all of our requirements.

The future of transistor technology will require a collaborative approach that brings together different disciplines and expertise. It will require a shift away from the traditional "general-purpose" approach to technology development toward application-specific solutions that are tailored to the specific needs of different sectors, such as AI, high-performance computing (HPC), and augmented reality/virtual reality.

In this article, we explore the challenges and opportunities facing the semiconductor industry over the next 75 years. Specifically, we examine the technologies that are likely to shape the future of computing, including the development of the next generation of CMOS transistors, new materials with unique properties, and new computing beyond the traditional von Neumann architecture. We also discuss the trend toward heterogeneous integration, where different types of chips are combined to create more powerful systems. In addition, we also discuss the importance of design technology

Digital Object Identifier 10.1109/MSSC.2023.3280433
Date of current version: 20 August 2023

cooptimization (DTCO) and system technology cooptimization (STCO) frameworks in addressing the major system scaling bottlenecks. Finally, we highlight the growing importance of sustainability in the semiconductor industry. As the industry continues to scale, it must adopt more sustainable practices to minimize its impact on the environment and ensure long-term viability.

While we cannot predict the future with certainty, we can be sure that the next 75 years will be filled with challenges and opportunities. As we navigate this transition, we must remember that collaboration and multidisciplinary expertise will be key to unlocking the full potential of the technologies that will shape our future.

Navigating the Scaling Walls: From FinFET to Nanosheets to Complementary FET

Over the next decades, transistors will remain essential components in many general-purpose computing applications. The substantial efficiencies achieved through years of optimizing transistor technology will continue to be leveraged. While new compute paradigms like quantum computing hold great promise for solving some of the most challenging computational problems, they are unlikely to replace general-purpose computers entirely.

When I started my Ph.D., there was a growing concern in the semiconductor industry that CMOS technology might be approaching a scaling limit. At the time, many researchers believed that CMOS devices could not be reliably scaled below 100 nm, due to the physical properties of the materials used to create them.

Even as we celebrate this milestone in semiconductor history, we find ourselves at a critical inflection point.

However, over the past decades, the industry has managed to continue scaling CMOS technology in a way that was not thought possible at the time. Through the advancements in EUV lithography and innovative approaches to materials and device architectures, such as the use of FinFETs [1], [2] and high-κ gate dielectrics [3], Moore's law has been able to continue, leading to significant increases in computational power and energy efficiency. Additionally, silicon-on-insulator (SOI) technologies, long contenders to FinFET, have also found their unique place in the road map, particularly for low-power and high-frequency applications [4].

Despite these advances, there are growing concerns that the industry is reaching the limits of what can be achieved through scaling CMOS. As feature sizes continue to shrink, researchers are encountering a range of physical and technical challenges that are becoming increasingly difficult to overcome.

The semiconductor industry has made the decision to transition from FinFET to nanosheets for the 3/2-nm technology node, with the forksheet being the next evolutionary step from nanosheets, as can be seen in Figure 1. This architecture enables a reduction in n-to-p spacing, which can lead to further area scaling by reducing the standard cell height. To extend the lifetime of transistor scaling well beyond the 1-nm node, the industry is exploring the use of complementary FET (CFET) [5], [6], where NMOS and PMOS devices are stacked on top of each other. This vertical 3D stacking allows for further optimization of effective channel width and drive current, while the resulting area scaling can be used to push digital standard cell heights to four tracks and below.

While, on paper, CFET makes a lot of sense, it pushes the limits of what materials and process tools can do today. Many processes and tools required for the development of these devices are still in the early stages of research and development, making it difficult to manufacture them at scale, and CFET must therefore be seen as a technology entering production more than a decade from now.

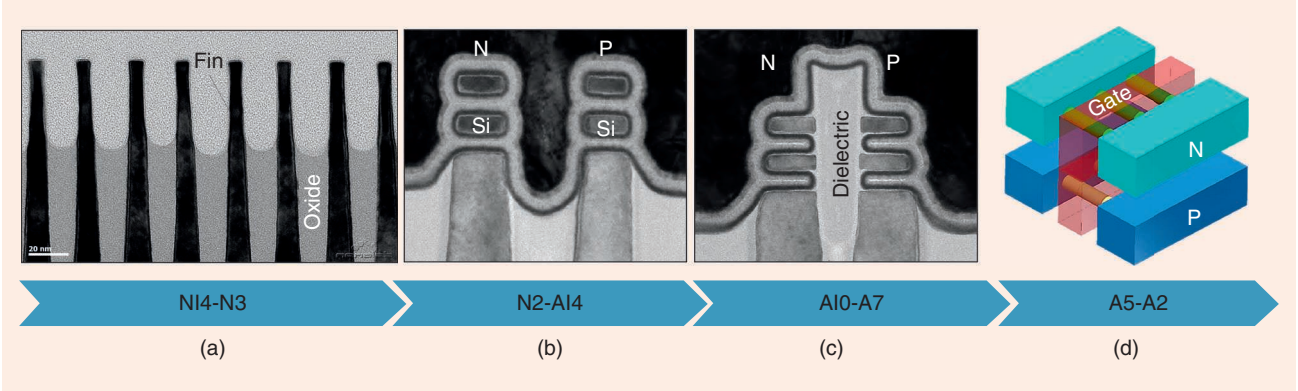


FIGURE 1: The logic transistor scaling road map: (a) FinFET, (b) nanosheets, (c) forksheets, and (d) complimentary FET. N14: 14nm node; A7: 7A node.

It is becoming increasingly clear that we cannot rely on a single technology to fulfill all of our requirements.

What does it mean to the design community? Already, with the introduction of FinFET, designers had to adapt to new design rules, methodologies, and tools. With the advent of CFET, designers will face a new set of challenges. CFET is a pure 3D device. This means that designers will need to consider new factors, such as vertical stacking and increased complexity in device connectivity. Another significant difference between CFET and FinFET technologies is that CFET will likely be a logic device only, meaning that implementing I/O devices and other functionalities, such as RF, will be difficult and require some form of heterogeneous integration. This may limit the use of CFET technology to specific applications, such as HPC or AI.

In conclusion, the semiconductor industry has managed to continue scaling CMOS technology in innovative ways to increase computational power and energy efficiency, but concerns remain that the industry is reaching the limits of what can be achieved through pure transistor scaling. The proposed CFET architecture offers a potential solution for extending the lifetime of pure transistor scaling beyond the 1-nm node, but it pushes the limits of current materials and process tools, and designers will face new challenges in adapting to this technology.

2D Materials: The Next Si? Or Something Else Entirely?

In addition to Si-based transistors, the future of electronics is likely to see the coexistence of a variety of new switch types based on different materials and physical principles. One such promising area is 2D materials, which have been receiving a lot of attention in recent years, due to their unique properties, such as high mobility, flexibility, and tunability. In particular, 2D metal dichalcogenides, carbon nanotubes, compound semiconductors, semiconducting oxides, and other yet-to-be-discovered semiconductors are being explored as channel materials for transistors, with the potential to overcome the limitations of Si-based transistors in terms of scalability and power efficiency.

A large part of the research today is focused on 2D materials. 2D monolayer crystalline materials have emerged as promising candidates for replacing Si in CMOS channels, due to their unique properties [7]. Materials such as tungsten disulfide and hafnium disulfide are examples of transition metal dichalcogenides that can be used as active layers in new transistors. The conduction in these channels occurs in atomic-thin layers, allowing for gate length scaling down to <10 nm with a high-κ gate stack, without significant short-channel effects.

Now, some might argue that we have witnessed a similar narrative in the past with the emergence of III-V materials as a potential replacement for Si in CMOS. However, there are reasons to believe that the situation with 2D materials may be different. For one, physics plays a lot more in favor of these materials. While III-V materials were found to lose their advantage over Si at scaled dimensions, a lower dielectric constant, higher bandgap, and high mobility at ultrathin dimensions favor the use of 2D materials over high-mobility materials like III-V and even Si. There is a fundamental limitation to how far Si can be scaled down. At dimensions of a few nanometers, the mobility in Si nanosheets decreases significantly, while 2D materials hold the promise of high mobility at these dimensions.

However, while the potential benefits of 2D channels are significant and substantial progress has been achieved in the past years, there are several remaining challenges in integrating these materials, including growth quality, which affects mobility and performance, formation of a good quality gate dielectric, doping, and contact resistance. When these materials finally find their way into commercial transistors, it is highly probable that they will be utilized in the form of a nanosheet or CFET architecture [8]. The ultrathin sheets of 2D materials are particularly suitable for stacking, thereby improving the transistor's drive current and reducing its footprint.

The future of computation is not limited to transistors alone, as new switches based on different physical principles are also being developed. Magnetic spin transfer, for instance, holds promise as a means of data storage and computing, while photonics [9] is being explored to transmit data at high speeds over long distances. Ferroelectricity [10] is another area of interest, with potential applications in low-power memory and logic devices. As these new types of switches are developed, it will be important for designers and engineers to have access to new electronic design automation (EDA) tools and design methodologies that can accommodate the different physics and material properties of these emerging devices.

Everything Heterogeneous

Heterogeneous integration, which involves blending different technologies together to create a more complex and advanced system, is becoming increasingly important in the semiconductor industry as the need for more sophisticated and specialized applications grows. Over the next 75 years, the use of heterogeneous integration is likely to play a crucial role in the development of new IC technologies.

Traditionally, the semiconductor industry has focused on scaling down device dimensions to achieve better performance, but this approach is reaching its limits due to various challenges, such as power consumption,

heat dissipation, and reliability. Heterogeneous integration offers a new path forward by combining traditionally dissimilar devices and materials to create more powerful and (cost-) efficient systems.

The chiplet approach has gained a lot of interest over the past years, very recently by the automotive industry. Chiplets are smaller specialized components that can be combined to create a larger more complex IC. One of the key advantages of chiplets is their ultimate flexibility, as each chiplet can be built in a completely different type of technology or technology node. They are connected using high-speed interfaces such as the unified interconnect standard, which provides low-latency high-bandwidth links between chiplets. But, as can be seen from Figure 2, chiplets are not the only option. There is a zoo of different technology features that can be used for enabling heterogeneous integration [11], [12], categorized according to 3D interconnect pitch capability, from package-level stacking all the way to transistor-level stacking.

For example, 3D die stacking using TSVs and solder bump stacking technology is used to produce compact mobile systems with power delivery and I/O functions on one chip and compute and graphics functions on another [13]. Memory and logic dies, separately designed and connected on active Si interposers, leverage ease of design and existing processes like die-to-die solder bump stacking.

The future of transistor technology will require a collaborative approach that brings together different disciplines and expertise.

However, determining when hybridization makes sense and choosing the best 3D technology for a particular application requires specialized codesign flows and EDA tools, and the industry is still in the early stages of this development.

STCO: The Multidisciplinary Engineer

This brings us to STCO. It is a framework for optimizing the performance of complex systems by cooptimizing the technology, architecture, and applications and thus goes hand in hand with the trend toward more heterogeneity [14].

The need for STCO arises from the fact that traditional methods of predicting the future of CMOS technologies, such as Moore's law and Dennard's scaling laws, are no longer reliable indicators of future performance. STCO represents a significant advancement beyond today's traditional DTCO. The DTCO methodology has gained importance over the past decade, especially with the introduction of the FinFET technology, and aims to achieve optimal design by balancing performance, power, and area requirements with manufacturing capabilities and limitations. Before DTCO, it was all about pure transistor dimensional scaling. In the DTCO era, the focus shifted toward track height scaling.

While new device architectures and scaling boosters, supported by DTCO, can maintain some area scaling in the next couple of CMOS nodes, they will not suffice in providing the system scaling expectations of future applications.

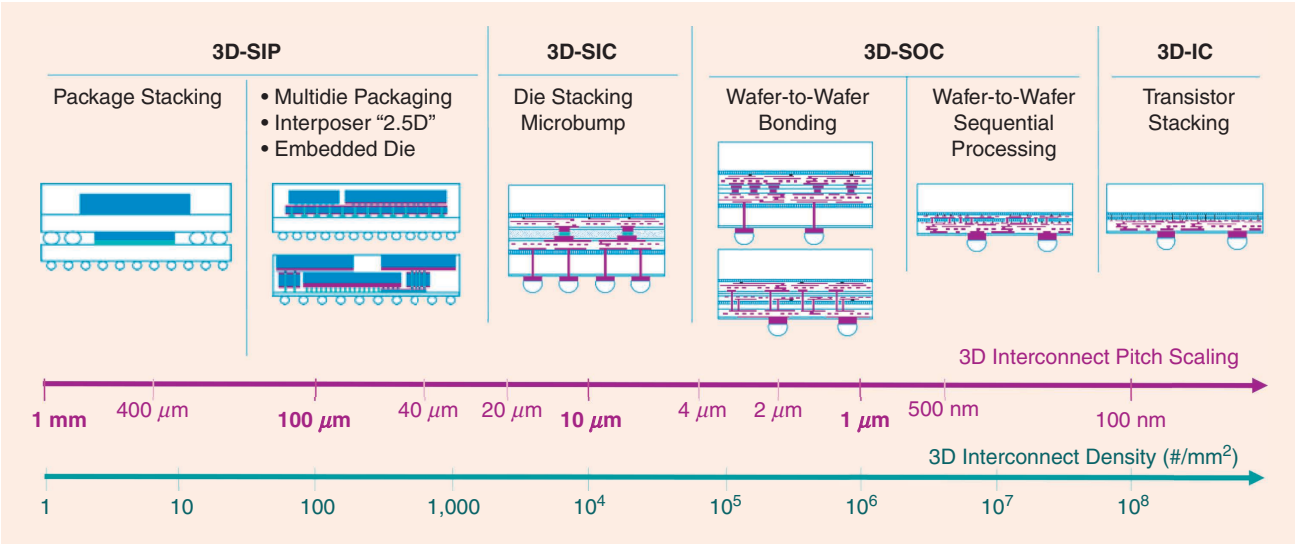


FIGURE 2: The landscape of 3D interconnect technologies can be categorized based on their increasing 3D interconnect density capability. System-in-package (SIP) integration technologies and stacked-IC (SIC) technology with die-to-die interconnects offer relatively lower-density capabilities. On the other hand, system-on-chip (SOC) integration enabled by advanced codesign EDA tools, wafer-to-wafer bonding technologies, and sequential 3D stacking offer higher-density capabilities. The highest density is offered by transistor-level stacking of devices, allowing for 3D integration at the standard cell level, which is referred to as *3D-IC* [11]. Reprinted, with permission, from [7] S. B. Samavedam et al., "Future logic scaling: towards Atomic Channels and Deconstructed Chips," *IEDM Tech. Digest*, 2020.

As the industry continues to scale, it must adopt more sustainable practices to minimize its impact on the environment and ensure long-term viability.

To address these challenges, a true system-architecture-to-technology optimization loop is required. An STCO framework can address the right technology ingredients that can unlock the major system scaling bottlenecks and enable new ways of computing. In addition, STCO can enrich the technology road map with system scaling challenges and application-specific requirements (Figure 3).

To achieve optimal system implementation, multiple technologies, such as extreme CMOS logic scaling, advanced 3D packaging, novel memory elements, and even Si photonics, must be optimized in a subtle manner. Previously, these different technology research activities have been conducted separately, each with their own DTCO research and road map. Hardware/software codesign will become ubiquitous.

In the next 75 years, STCO will play a critical role in the semiconductor industry's evolution. Solutions will be driven by application-specific requirements, and optimal system implementation will require a subtle optimization across multiple technologies.

The move toward STCO represents a significant shift in the way that engineers and researchers need to approach the semiconductor industry. Traditionally, engineers were trained in one particular area of expertise, such as materials science, electrical engineering, or computer science. However, to truly optimize systems and technologies for future applications, engineers and researchers will need to develop a more multidisciplinary skill set that spans multiple areas of expertise.

This means that the next generation of engineers and researchers will need to be comfortable working across different disciplines and collaborating with colleagues from a variety of backgrounds. They will need to be skilled in areas such as materials science, electrical engineering, computer science, even biology, and many more.

To prepare for this future, universities and other educational institutions will need to rethink their engineering curricula to ensure that they are providing students with the necessary interdisciplinary skills and knowledge. This may involve developing new programs or courses that focus on STCO or incorporating STCO principles into existing courses. Ultimately, the success of STCO will depend on the ability of engineers and researchers to work together across disciplines and leverage their combined expertise to develop optimized systems and technologies for the future.

Beyond von Neumann

It is hard to imagine the future of IC technologies without considering the limitations of the von Neumann architecture, which consists of a processing unit, memory, and I/O devices. The processing unit contains the CPU, which executes instructions and performs arithmetic and logic operations. Memory stores both instructions and data that the CPU uses for processing, and I/O devices allow the computer to communicate with the outside world.

While this architecture has been the foundation of computing for decades, it is facing challenges in the era of big data and AI. The traditional von Neumann architecture, where the CPU and memory are separate, creates a bottleneck in the processing

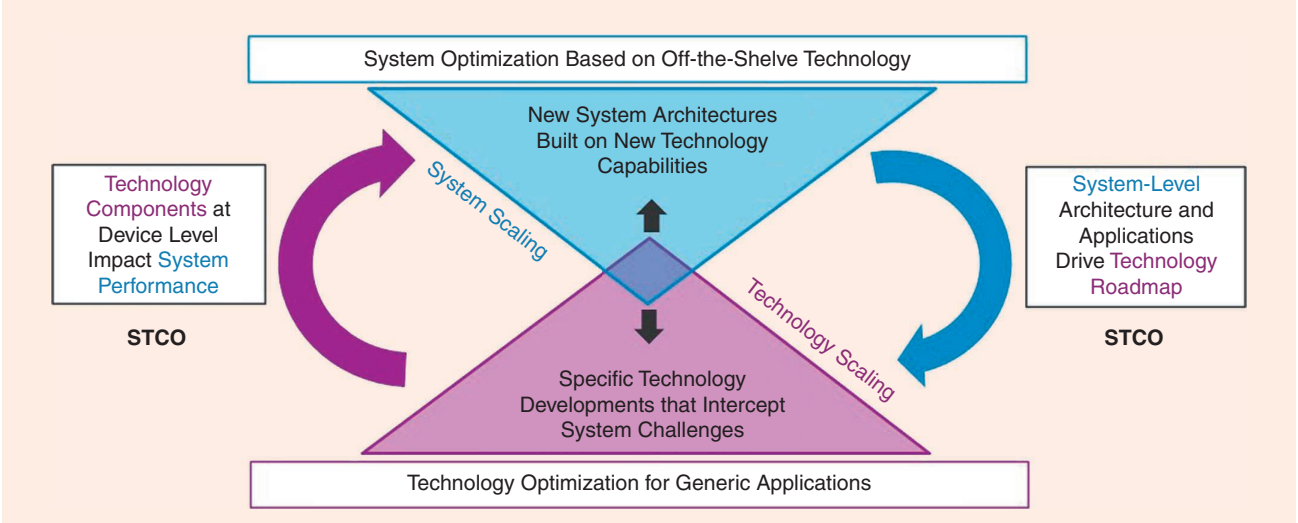


FIGURE 3: The STCO framework.

of large datasets. Additionally, it requires significant energy to move data between the two, limiting the efficiency of the system.

To address this memory wall, research has been focusing on, among others, in-memory computing. By storing and processing data in the same location, in-memory computing can significantly improve the performance and energy efficiency of computing systems, especially in applications that require frequent and fast data access. Various memory technologies have been considered for in-memory computing, including resistive RAM (RRAM), phase change memory (PCM), and magnetic RAM (MRAM). While each of these technologies offers advantages, such as fast access, high capacity, and nonvolatile storage, they also have their own challenges. For example, RRAM is still in the early stages of development, and researchers are working to address challenges in achieving high yields and reliable operation. PCM offers nonvolatile storage and fast access times, but it has relatively low endurance compared to other memory technologies. MRAM also offers fast access times and nonvolatile storage but currently has lower densities and higher power consumption. The ultimate memory technology to address all these requirements is still a topic of extensive research that, likely in the coming years and decades, will see significant breakthroughs.

In addition to these challenges, there is a need to develop algorithms and architectures that can fully exploit the capabilities of in-memory computing.

Besides this, other architectures are being explored to overcome these limitations, such as neuromorphic and quantum computing. Neuromorphic computing utilizes a massively parallel approach, similar to the way the human brain processes information. Quantum computing, on the other hand, uses quantum bits to perform calculations that would be impossible with classical computers.

The future of electronics is likely to see the coexistence of a variety of new switch types based on different materials and physical principles.

In addition to these alternative architectures, AI is playing an increasingly important role in the development of next-generation circuits. Machine learning algorithms, for example, can optimize the design of circuits and systems. AI is also being used to improve the performance and reliability of IC manufacturing processes, reducing defects, and increasing yield.

As technologies continue to evolve, a multidisciplinary approach will be needed to address the complex challenges that lie ahead. Beyond von Neumann, we must explore new computing architectures, integrate diverse technologies through heterogeneous integration, and leverage the power of AI to optimize designs and manufacturing processes. The next 75 years will undoubtedly see tremendous advances in all areas, and the engineers and researchers of the future will need to be adaptable and innovative to stay at the forefront of this rapidly evolving field.

Sustainability in the Semiconductor Industry

The IC manufacturing processes have a significant impact on the environment, and as demand for ICs continues to grow, so too does the need for sustainable solutions [15]. The information technology sector is estimated to be responsible for up to 5% of the total global emissions in terms of equivalent carbon dioxide [16]. This includes the impact of all types of devices, from data centers to mobile phones, as well as their use and manufacturing and the energy used by the networks. Although only a fraction of this impact can be attributed to semiconductor manufacturing, it remains a resource-intensive process with a complex supply chain that consumes significant amounts of energy, water, chemicals, and raw materials.

While the life cycle assessment (LCA) is a well-established methodology for evaluating the environmental impact of products, conducting a good LCA for ICs is still challenging due to the scarcity of information. Most data today are outdated. As a result, errors and variability occur in the LCA of electronic products, making it difficult to accurately assess their environmental impact.

To address this issue, there is a pressing need to close the data gap in semiconductor fabrication. This involves creating a quantified view of energy consumption, water usage, mineral usage, and greenhouse gas emissions associated with manufacturing processes. By obtaining accurate and up-to-date data, the semiconductor industry can gain a better understanding of the environmental impact of its products and take steps to minimize its carbon footprint.

Closing that data gap requires collaboration across the industry, with stakeholders working together to gather and share data on semiconductor manufacturing processes. This can involve developing new measurement techniques, improving data collection, sharing practices, and establishing common standards and frameworks for reporting environmental impacts. Ultimately, this will help to promote transparency, accountability, and innovation while also ensuring that the industry is doing its part to address the urgent environmental challenges of our time.

However, to achieve sustainable computing, it is essential to consider both the embodied (manufacturing) and operational carbon footprints [17]. This calls for the development and

implementation of robust carbon metrics and a design framework that encompasses the entire life cycle of cloud computing. By quantifying and optimizing the embodied carbon emissions along with the operational energy consumption, it becomes possible to make informed decisions regarding hardware design, energy efficiency, and the adoption of renewable energy sources.

In addition to environmental impact, sustainable practices should also encompass social and economic impact. The semiconductor industry must prioritize employee well-being by ensuring a safe and healthy work environment, offering flexible work arrangements, and supporting mental health and well-being initiatives. As the industry continues to evolve and change over the next 75 years, it is important to note that a new generation of researchers and engineers with a heightened awareness of the impact of their work on both the environment and society will emerge. It is essential for the industry to embrace and support this new generation.

Overall, the next decades will be a critical period for the semiconductor industry to prioritize sustainability in all aspects of its operations.

Conclusions

The next 75 years of IC technologies will be characterized by a range of challenges and opportunities. While the traditional Moore's law road map may no longer be valid, the industry is entering an exciting new phase of innovation and growth. To succeed in this environment, designers will need to be familiar with a range of different technologies and fields, from materials science and device physics to computer science and AI. They will need to be able to collaborate effectively with experts from these fields as well as understand the various tradeoffs that exist between different technologies and design choices. With the right mindset and skill set, designers can help to shape the future of semiconductor technologies, creating products that are not only technically sound but also socially responsible and environmentally sustainable.

Acknowledgment

The author would like to thank S. Samavedam, J. Ryckaert, J. Myers, G. Hellings, N. Horiguchi, and G. Sankar Kar for their invaluable insights and constructive discussions. The author would also like to express gratitude to the partner companies that support the Imec industrial affiliation programs and to all the diligent researchers at Imec whose work has contributed to this article.

References

[1] D. Hisamoto et al., "A folded-channel MOSFET for deep-sub-tenth micron era," in *Proc. Int. Electron Devices Meeting (IEDM), Tech. Dig. (Cat. No.98CH36217)*, 1998, pp. 1032–1034, doi: 10.1109/IEDM.1998.746531.

[2] J. Kavalieros et al., "Tri-gate transistor architecture with high-k gate dielectrics, metal gates and strain engineering," in *Proc. Symp. VLSI Technol., Dig. Tech. Papers*, 2006, pp. 50–51, doi: 10.1109/VLSIT.2006.1705211.

[3] A. Chatterjee et al., "CMOS metal replacement gate transistors using tantalum pentoxide gate insulator," in *Proc. Int. Electron Devices Meeting, Tech. Dig. (Cat. No.98CH36217)*, 1998, pp. 777–780, doi: 10.1109/IEDM.1998.746471.

[4] S. N. Ong et al., "A 22nm FDSOI technology optimized for RF/mmWave applications," in *Proc. IEEE Radio Freq. Integr. Circuits Symp. (RFIC)*, 2018, pp. 72–75, doi: 10.1109/RFIC.2018.8429035.

[5] J. Ryckaert et al., "The complementary FET (CFET) for CMOS scaling beyond N3," in *Proc. IEEE Symp. VLSI Technol.*, 2018, pp. 141–142, doi: 10.1109/VLSIT.2018.8510618.

[6] N. Horiguchi and J. Ryckaert. "Integrating CFET into the logic technology roadmap beyond 1 nm." Embedded. Accessed: Aug. 1, 2022. [Online]. Available: https://www.embedded.com/integrating-cfet-into-the-logic-technology-roadmap-beyond-1-nm/

[7] S. B. Samavedam et al., "Future logic scaling: Towards atomic channels and deconstructed chips," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2020, pp. 1.1.1–1.1.10, doi: 10.1109/IEDM13553.2020.9372023.

[8] Y.-Y. Chung et al., "First demonstration of GAA monolayer-MoS2 nanosheet nFET with 410μA μ m ID 1V VD at 40nm gate length," in *Proc. Int. Electron Devices Meeting (IEDM)*, 2022, pp. 34.5.1–34.5.4, doi: 10.1109/IEDM45625.2022.10019563.

[9] J. Van Campenhout, "Silicon photonics technology for terabit-scale optical I/O," in *Proc. ACM/IEEE Int. Workshop Syst. Level Interconnect Prediction (SLIP)*, 2021, p. 41, doi: 10.1109/SLIP52707.2021.00015.

[10] S. Clima et al., "Dielectric response in ferroelectrics near polarization switching: Analytical calculations, first-principles modeling, and experimental verification," *IEEE Trans. Electron Devices*, vol. 69, no. 9, pp. 5345–5350, Sep. 2022, doi: 10.1109/TED.2022.3194822.

[11] E. Beyne, "The 3-D interconnect technology landscape," *IEEE Des. Test*, vol. 33, no. 3, pp. 8–20, Jun. 2016, doi: 10.1109/MDAT.2016.2544837.

[12] E. Beyne, "Heterogeneous system partitioning and the 3D interconnect technology landscape," in *Proc. Symp. VLSI Technol. Circuits, Short Course SC2-2*, 2020.

[13] S. Khushu and W. Gomes, "Lakefield: Hybrid cores in 3D package," in *Proc. IEEE Hot Chips 31 Symp. (HCS)*, 2019, pp. 1–20, doi: 10.1109/HOTCHIPS.2019.8875641.

[14] "Unlocking system-scaling bottlenecks with STCO," interview Julien Ryckaert. [Online]. Available: https://www.imec-int.com/en/articles/unlocking-system-scaling-bottlenecks-system-technology-co-optimization

[15] M. G. Bardon, "DTCO including sustainability: Power-performance-area-cost-environmental score (PPACE) analysis for logic technologies," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2020, pp. 41.4.1–41.4.4, doi: 10.1109/IEDM13553.2020.9372004.

[16] U. Gupta et al., "Chasing carbon: The elusive environmental footprint of computing," in *Proc. IEEE Int. Symp. High-Perform. Comput. Archit. (HPCA)*, 2021, pp. 854–867, doi: 10.1109/HPCA51647.2021.00076.

[17] C.-J. Wu, "Carbon impact of computing," in *Proc. Int. Solid-State Circuits Conf. (ISSCC) Panel Sustainability*, 2023.

About the Author

Nadine Collaert (collaert@imec.be) is a program director at Imec, 3001 Heverlee, Belgium. She is currently responsible for the advanced RF program looking at heterogeneous integration of III-V/III-N devices with advanced CMOS to tackle the challenges of next-generation mobile communication. Before that, she was program director of the LOGIC Beyond Si program focused on the research of novel CMOS devices and new material-enabled device and system approaches to increase functionality. She has been involved in the theory, design, and technology of FinFET devices, emerging memories, transducers for biomedical applications, and the integration and characterization of biocompatible materials. She has a Ph.D. in electrical engineering from KU Leuven, and she has more than 400 publications and holds more than 10 patents in the field of device design and process technology. She is a Member of IEEE.

SSC