

Beyond-CMOS VLSI (Future Technology after Silicon)

Ammad Mallick¹, Farhan Ahmed², M Zamin Ali Khan³, Fatima Ilyas⁴, Faigha Karim⁵,Khalid Bin Muhammad⁶, Muhammad Usama Khan⁷, Hassan Iqbal⁸¹Department of Computer Science, Cardiff Metropolitan University, London, UK²Researcher, Crystallite Pvt Ltd, Karachi, Pakistan.³Professor/HOD Department of Computer Science, Iqra University, Pakistan

Correspondence Email: muhammad.zamin@iqra.edu.pk

⁴Researcher, Habib Metro, Karachi, Pakistan⁵Engineer, Esr Tech, Pakistan⁶Associate Professor, Department Of Computer Science, Iobm,, Pakistan⁷Development Manager, Uhf Solutions, Pakistan⁸Researcher, Charles Strut University, Melbourne Campus, Australia*Abstract*

The scaling of traditional CMOS technology based on silicon is rapidly reaching physical and economical limits for deepscaled nodes below 5 nm, as has been expounded in various recent works on scaling options [1], [3]. The presence of short-channel effects, high power density, variability, and interconnect limitations has made the pursuit of Beyond-CMOS device options imperative, as has already been discussed for sub-5 nm nodes [1], [4]. This paper seeks to provide a comprehensive comparison and assessment of recent efforts on advanced CMOS technologies, two-dimensional material transistors, Tunnel Field Effect Transistors, and two-dimensional magnetic heterostructures for spintronics and quantum computing, as has appeared in various recent surveys on the topic. Through a validation of scalability, possible trade-offs for power and performance, manufacturability, and CMOS integration, the paper argues that a heterogeneous integration strategy that packages more than one device technology holds the most promise for maintaining Moore's scaling momentum [1], [15]. Notably, the need for such scaling is highly relevant for modern VLSI systems, because scaling directly affects the reliability, energy efficiency, yields, and overall system performance, as has appeared consistently in the literature on the topic [1], [18]. Moreover, with further scaling, previously sub-dominant physical phenomena are increasingly governing device operations, necessitating innovations at both the materials and architectural levels [3], [4].

Keywords: Beyond-CMOS, VLSI Scaling, Two-Dimensional Materials, TFET, Spintronics, 2D Magnetic Heterostructures, Quantum Computing

I. INTRODUCTION

For over five decades, Moore's Law has been driving exponential growth in computing capability through the incessant scaling of CMOS technology. However, as the dimensions approach atomic length scales, conventional scaling is facing severe limitations due to quantum tunneling, degradation in electrostatics, thermal management issues, and rising complexity in fabrication, documented in advanced CMOS scaling reports. Although innovations in FinFETs and GAA nanosheet transistors extended the viability of CMOS, they were not sufficient as standalone long-term solutions. Hence, an increasing amount of effort has been done in researching Beyond-CMOS technologies with the ability to complement or augment conventional CMOS platforms. This fact is of critical importance for modern VLSI systems because further miniaturization directly impacts device reliability, power efficiency, fabrication yield, and overall system performance, reflected in several CMOS roadmap studies. As the scaling continues, physical effects that have been negligible at larger scales now dominate device behavior, requiring material-level and architectural innovation to achieve continued progress in computing capability.

II. CONVENTIONAL CMOS SCALING: PRESENT SITUATION AND UPCOMING BOUNDARIES

A. CMOS Scaling Evolution

The main motivation of conventional silicon-based CMOS technology, the basis for VLSI systems for over 50 years, has been coupled by Moore's Law and Dennard scaling principles [2, 3]. Constant decreases in gate length, oxide thickness, and supply voltage allowed for exponential increases in transistor density and performance while managing power dissipation [1], [18]. However, higher leakage currents, unpredictability, and parasitic effects caused the

classical scaling laws to break down as device dimensions moved into the deep-submicron and nanometer regimes [1], [4].

B. Sophisticated CMOS Structures

The industry switched from planar MOSFETs to FinFETs and, more recently, gate-all-around (GAA) nanosheet and nanowire transistors to reduce short-channel effects at advanced nodes [3], [17]. Scaling below 7 nm is made possible by these multigate topologies, which enhance electrostatic control over the channel and lessen drain-induced barrier lowering (DIBL) [1], [18]. Despite these advancements, overall device performance is increasingly dominated by parasitic resistance, contact scaling, and interconnect latency [3], [16].

C. Immediate Economic and Physical Boundaries

Quantum mechanical factors as mobility deterioration, gate leakage, and direct source-to-drain tunnelling have a major impact on device reliability and power efficiency at technology nodes smaller than 5 nm [1], [4]. Moreover, the economic viability of additional CMOS scaling is threatened by extreme ultraviolet (EUV) lithography, intricate process integration, and growing fabrication prices [1], [18]. Instead of completely replacing CMOS, these limitations encourage the investigation of other device designs that can enhance or expand it [6], [15]. As seen in Fig. 1, conventional CMOS scaling has progressed from planar MOSFETs to FinFETs and gate-all-around (GAA) architectures to eliminate short-channel effects and enable sub-5 nm technology nodes.

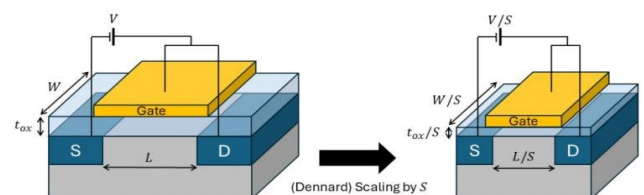


Fig. 1: Scaling of CMOS and its Issues

III. SEMICONDUCTOR TRANSISTORS IN TWO DIMENSIONS (2D)

A. Reasons to Use 2D Channel Materials

In light of their atomically small body thickness and superior electrostatic characteristics, two-dimensional semiconductor materials including MoS_2 , WS_2 , WSe_2 , and black phosphorus have emerged as intriguing candidates for Beyond-CMOS transistors [5], [6]. In contrast to bulk silicon, 2D materials inherently minimize short-channel effects, allowing for aggressive scaling without significant leakage penalties [2], [7].

B. Advantages of Device Physics and Performance

High-quality gate dielectrics with lower interface trap density are made possible by the lack of dangling bonds at van der Waals interfaces, which improves subthreshold slope and lowers variability [5], [6]. Even at sub-10 nm channel lengths, experimental demonstrations have demonstrated reliable switching behavior and high on/off current ratios [7], [8]. For low-power and high-density logic applications, these features make 2D FETs very appealing [4], [6].

C. Manufacturing and Integration Difficulties

Despite their benefits, 2D transistors have a lot of obstacles to overcome before they can be widely used. It is still challenging to achieve consistent, wafer-scale synthesis utilizing CMOS-compatible methods [6], [7]. Performance reproducibility is further restricted by contact resistance, doping control, and variability across wide areas [5], [8].

To enable practical VLSI implementation, ongoing research focusses on hybrid integration with silicon platforms and direct growth approaches [6], [15]. As illustrated in Fig. 2, two-

dimensional (2D) semiconductor transistors use atomically thin channel materials like MoS_2 and WS_2 to enable better electrostatic control, less short-channel effects, and improved scalability at advanced technology nodes.

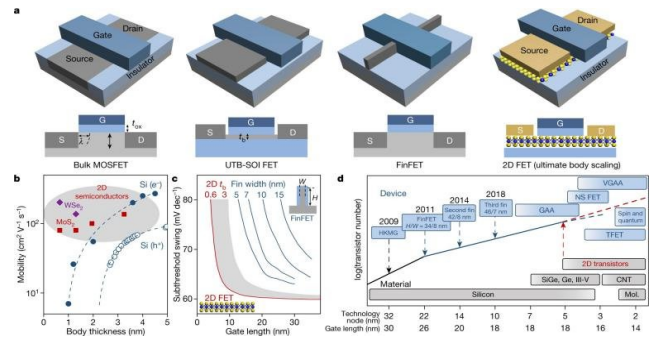


Fig. 2: Promises and prospects of 2D transistor

IV. FIELD-EFFECT TRANSISTORS IN TUNNELS (TFETs)

A. Principle of Operation

Instead of using thermionic emission, Tunnel Field-Effect Transistors (TFETs) use band-to-band tunnelling (BTBT) to enable sub-60 mV/decade subthreshold swing at ambient temperature [8], [10]. Because of this feature, TFETs may operate at much lower supply voltages than traditional MOSFETs, which directly lowers dynamic power consumption [6], [9].

B. Device Architectures and Materials

Device structure and material choice have a significant impact on TFET performance.

Materials with narrow bandgaps, such as germanium, 2D semiconductors, and III–V compounds, increase the likelihood of tunnelling and improve on-current [6], [10]. To maximize electric field distribution and tunnelling efficiency, sophisticated topologies such as heterojunction TFETs, vertical TFETs, and dual-gate configurations have been proposed [8], [11].

C. Benefits and Drawbacks

TFETs are ideal for ultra-low-power applications such near-threshold computer systems, biomedical implants, and Internet of Things devices [6], [8]. However, their use in high-performance logic is hampered by restricted drive current, sensitivity to interface flaws, and manufacturing complexity [3], [9]. Thus, hybrid integration with CMOS is thought to be the most realistic approach in the near future [1], [6]. As seen in Fig. 3. Tunnel Field-Effect Transistors (TFETs) enable ultra-low-voltage operation with sub-60 mV/decade subthreshold swing through band-to-band tunnelling at the source-channel junction.

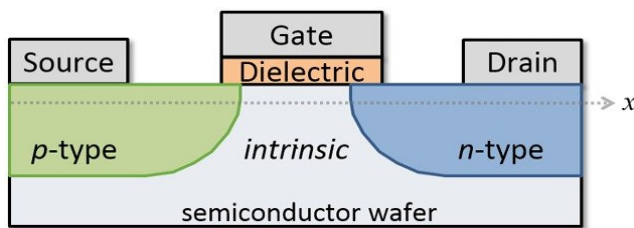


Fig. 3: Tunnel Field Effect Transistor

V. SPINTRONICS AND 2D MAGNETIC HETEROSTRUCTURE

A. The paradigm of spin-based computing

Fundamentally new device functions like non-volatility and ultra-low standby power are made possible by spintronics, which takes advantage of the spin degree of freedom of electrons rather than charge [11], [21]. Because of their intrinsic magnetism at atomic thicknesses, two-dimensional magnetic materials such as CrI_3 , Fe_3GeTe_2 , and MnBi_2Te_4 can be integrated at the nanoscale [12], [13].

B. Concepts and Uses of Devices

Spin-orbit torque switching, proximity-induced magnetism, and magneto resistive effects all essential for non-volatile memory, logic-in-

memory, and neuromorphic computing architectures are made possible by 2D magnetic heterostructures [11], [14]. The von Neumann bottleneck present in traditional CMOS systems is overcome by these devices' ability to store and process data simultaneously [14], [15].

C. Difficulties and Prospects

Achieving reliable room-temperature magnetic ordering, scalable manufacturing, and smooth integration with CMOS back-end procedures are major obstacles [12], [13]. Despite these obstacles, spintronic devices are positioned as an essential part of future Beyond-CMOS VLSI systems because they provide long-term potential for energy-efficient and quantum-assisted computing platforms [11], [21]. Two-dimensional magnetic heterostructures offer non-volatile and energy-efficient device operation by using electron spin rather than charge to enable spin-based information processing, as shown in Fig. 4.

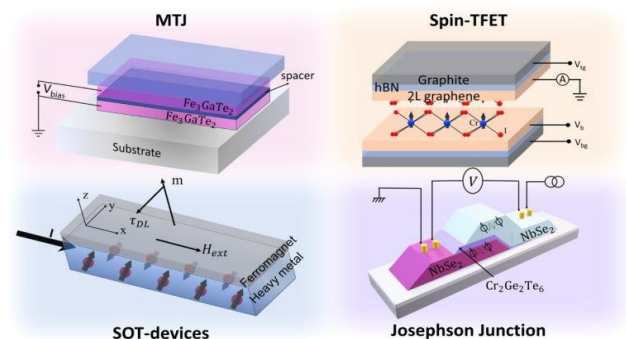


Fig. 4: 2D Magnetic Heterostructures and Spintronics

VI. REVIEW METHODOLOGY

The current literature review is conducted using a qualitative comparative approach based on peer-reviewed literature reviewed over the last five years. In cases where foundational literature has been published, these pieces of literature are used

for theoretical foundation purposes [1] and [3]. The different technologies reviewed are examined for critical parameters such as electrostatic scalability, power efficiency, compatibility, maturity of technology, and overall impact at a systems level, in accordance with past comparative studies previously documented in existing literature [6] and [11]. The assessment of these parameters assumes critical importance in existing VLSI technology as they affect miniaturization and consequently impact device reliability, efficiency, and overall system performance, as documented in past technology reports [1] and [18].

Moreover, there is a focus on trying to elucidate the general performance metrics, scalability trends, and typical technological bottlenecks for varying device types based on the associated platforms. In this case, particular attention is directed at the impact of material properties and device size with respect to variations in performance as reported within various studies. In a similar fashion, there is a focus on generalizing performance variations based on the size of the proof-of-concept implementation and

technologies near the threshold of industrial feasibility.

In addition, the system-level implications of innovations such as circuit compatibility and potential effects on the existing VLSI design flows are discussed at a qualitative level and serve the purpose of filling the innovation gap at the device level and implementation at the system level. While innovation at the scaling level advances, various physical phenomena that were until then insignificant become more critical at the device level and hence parallel innovation at the material and architectural levels is required along with the evaluation methodologies that view the system from the device level and circuit level simultaneously [4], [15].

VII. COMPARATIVE ANALYSIS OF BEYOND-CMOS TECHNOLOGY

Table I provides a comparative overview of traditional CMOS scaling and key Beyond-CMOS technologies, such as 2D semiconductors, TFETs, and spintronic devices.

Feature	CMOS Scaling	2D Semiconductors	Spintronics	TFETs
Maturity	Mature, industry-standard	Early-stage R&D	Research-stage, experimental	Early research
Scalability	Limited (<5 nm)	Atomic-scale, high density	Very high potential; 3D stacking possible	Moderate, material-dependent
Power Efficiency	Moderate; leakage issues	Low-power, reduced leakage	Ultra-low power; non-volatile	Ultra-low power, low voltage
Device Performance	Good but degrading at small nodes	High mobility, stable switching	Multifunctional; material-dependent	Low ON-current, high SS control
Manufacturability	Well-established fabs	Integration & wafer uniformity challenges	Experimental; production-scale challenging	Challenging, integration issues

Reliability	Established	Needs further validation	Thermal & interface stability concerns	Needs validation
Applications	Standard logic/memory	Next-gen ICs, IoT, AI	Neuromorphic, quantum, hybrid logic-memory	Low-power IoT, sensors, hybrid circuits

Table. I: Difference between CMOS scaling, 2D Semiconductor, Spintronics and TFET's

VIII. DISCUSSION

Comparative analysis across multiple technology platforms indicates that no single Beyond-CMOS device can independently replace conventional CMOS across all application domains without significant trade-offs [1], [6]. The most common high-performance CMOS architectures present in high-performance logic applications are advanced CMOS devices, such as FinFET and gate-all-around nanosheet transistors, both of which have

been matured to be suitable for production and densely packable into high density. Advanced CMOS devices such as FinFET devices and gate-all-around nanosheet devices are currently the dominant devices for high-performance logic applications because they are highly

manufacturable, yield good yields, and have a well-established design environment [3], [17]. As

scaling comes close to fundamental physical boundaries however, CMOS devices have been plagued by leakage power, variation, and by predominantly interconnect-limited performance constraints, which have curbed any further gains in energy efficiency and density.

Other emerging device concepts, such as Tunnel Field-Effect Transistors (TFETs) and two-dimensional (2D) semiconductor transistors, hold great promise as a solution to these electro-static and power scaling problems. The 2D material is promising for energy constrained device applications because, as a thin channel, 2D material offers excellent electrostatic control as 2D material, which makes it possible to enable very aggressive scaling and reduce short channel

effects [5] [8] while the TFETs overcoming the limitations of subthreshold swing can operate at extremely low voltage, attractive for energy constrained devices application. However, both technologies are still suffering from the problems of limited drive current, contact resistance, variant and manufacturability for larger scale use in the normal high-performance computing applications.

Alongside CMOS-based devices, spintronic devices offer additional support for the Beyond-CMOS world such as non-volatility, logic-in-memory and novel computing models not possible with charge-based CMOS devices [11, 15]. These features are important to reducing data transfer and standby power consumption, critical for data-heavy workloads, for emerging architectures based on quantum computing, and for the context of neuromorphic computing. The present spintronics devices are mainly experimental devices, however, the challenges of room-temperature operations, scalability production of spintronic devices and integration of spintronics devices with CMOS back-end are still to be resolved.

All these observations favor the interpretation that integration of a new system by heterogeneous mixing rather than complete replacement of one technology with another is the most expedient and promising strategy to future VLSI systems.

Combining advanced CMOS and selected Beyond-CMOS devices thereby allows designers to benefit from the best of both worlds and offset any weaknesses of either technology. As discussed in the system-level integration studies

[1] and [18] this is a critical aspect in modern VLSI systems as miniaturization of devices has direct impact on device level reliability, power consumption, yield, thermal management, and device level functionality.

Whereas earlier physical effects were inconsequential, in the modern chip they have become more important such as quantum tunneling, interconnect delay, reliability reduction due to process variability, and so on. The solution to these challenges calls for architectural, circuit and system integration level coordinated innovation along with material and device innovations. Hence, future advancement of VLSI technology will hinge on system-aware heterogeneous design methodologies that enable techniques to integrate different designs and device types to achieve performance, power, manufacturability, and long-term reliability benefits, beyond just the existing relationship of scaling a CMOS device [4, 21].

IX. LIMITATION AND CHALLENGES

First, experimental results are constantly changing, and second of all, because the various Beyond-CMOS device platforms are so young, and they use various benchmarking methods, it is challenging to compare the performance of different technologies directly. [3], [11] The often-smaller scale prototypes used in demonstrated prototypes are rarely tested under the same conditions, which restricts generalizing the outcomes.

Moreover, most Beyond-CMOS technologies are lacking in terms of long-term reliability, device-to-device variability, fabrication yields or manufacturability data for the technology, as noted in previous evaluations [1, 6]. These gaps make it difficult to evaluate Technology Readiness levels accurately, and create a huge difficulty to adopt at the circuit and system levels. Given the current trends in VLSI miniaturization,

this over-riding requirement is crucial from the point of view of device reliability, degree of power consumption, yield during fabrication, thermal stability and overall system performance of the VLSI system, and as a result it is imperative that experiments are carried out and statistically meaningful data recorded [1, 18].

The physical effects, which were inconsequential in the past, such as quantum confinement, tunneling leakage, interfacial defects and process induced variability, now dominate the device operation and performance metrics as scaling progresses. To meet these needs, progress at the material level and at the architectural level is required along with the establishment of standard evaluation methods to compete and consider technology selection for future VLSI based systems [15], [21].

X. CONCLUSION

Rather, as CMOS scaling moves slower and slower, it is time to find a new paradigm to break away from a single replacement technology and move into multiple complementary device concepts forming a number of heterodox architectures [1] [3]. Substituting CMOS-based architectures with other new technological components like two-dimensional (2D) semiconductor devices, Tunnel Field-Effect Transistors (TFETs) and spintronic devices is not an ideal way to replace it completely, instead it allows for a strategic co-integration of advanced CMOS architectures with Beyond-CMOS technologies, which offers the most credible and viable route for scaling performance, energy efficiency and enabling functional diversification for future VLSI systems [6], [11]. [22] [23]

Such heterogeneous integration can enable system designers to take advantage of the maturity, high yield, and performance of CMOS components for critical logic functions, as well as the ultra-low power operation of TFETs, the much more

superior electrostatic scalability of 2D materials, and spintronic components' non-volatility and logic-in-memory characteristics. As reported in semiconductor roadmaps and integration studies [1], [18], this aspect is particularly significant for modern VLSI systems as miniaturization further affects the reliability of the devices, power consumption and utilization, manufacturing yields, thermal handling and system level performance.

Quantum tunneling, process induced variation, interconnect delay and reliability degradation to name a few, become now the leading physical effects as scaling progresses, which dictate the system behavior and efficiency. At the level of materials or devices, innovations are needed to solve the challenges listed above, but also at the level of architecture and system integration. Thus, continued reliance on CMOS scaling alone is destined for the dustbin of history and the future will rely on strategies for heterogeneous integration that focus on achieving performance, power, manufacturability, and long-term reliability together across different devices and technologies [15, 21] rather than on individual scaling.

XI. REFERENCES

- [1] H. H. Radamson et al., "CMOS Scaling for the 5 nm Node and Beyond: Device, Process and Technology," *Nanomaterials*, vol. 14, no. 3, 2024.
- [2] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, 1965.
- [3] C. Auth et al., "A 10nm High Performance and Low-Power CMOS Technology," *IEDM Tech. Dig.*, 2017.
- [4] J.-P. Colinge et al., "Nanowire transistors without junctions," *Nature Nanotechnology*, vol. 5, 2010.
- [5] B. Radisavljevic et al., "Single-layer MoS₂ transistors," *Nature Nanotechnology*, vol. 6, 2011.
- [6] Shiraz Afzal, M.Zamin Khan et al "A Proposed OEIC Circuit with Two Metal Layer Silicon Waveguide and Low Power Photonic Receiver Circuit" Vol 9, Issue 5, pg 355-358, 2012
- [7] K. Kang et al., "Wafer-scale synthesis of monolayer MoS₂," *Nature*, vol. 520, 2015.
- [8] W. Y. Choi et al., "Tunneling field-effect transistors," *IEEE EDL*, vol. 28, 2007.
- [9] S. Salahuddin and S. Datta, "Use of negative capacitance," *Nano Letters*, vol. 8, 2008.
- [10] D. Sarkar et al., "A subthermionic tunneling transistor," *Nature*, vol. 526, 2015.
- [11] B. Zhang et al., "2D Magnetic Heterostructures: Spintronics and Quantum Future," *npj Spintronics*, vol. 2, 2024.
- [12] Y. Deng et al., "Gate-tunable room-temperature ferromagnetism," *Nature*, vol. 563, 2018.
- [13] S. Gong et al., "Discovery of intrinsic ferromagnetism in 2D materials," *Nature*, vol. 546, 2017.
- [14] J. Grollier et al., "Neuromorphic spintronics," *Nature Electronics*, vol. 3, 2020.
- [15] S. Datta, "Quantum Transport," *Cambridge Univ. Press*, 2005.
- [16] M. L. Lee et al., "Strained silicon channels," *IEEE EDL*, vol. 26, 2005.

[17] Q. Zhang et al., "Gate-all-around nanosheet CMOS," IEDM, 2022.

[18] I. Ferain et al., "Multigate transistors," Nature, vol. 479, 2011.

[19] T. Mikolajick et al., "Ferroelectric HfO₂ for memory," Advanced Materials, vol. 33, 2021.

[20] J. Appenzeller et al., "Carbon nanotube FETs," Phys. Rev. Lett., vol. 93, 2004.

[21] S. Wolf et al., "Spintronics: A spin-based electronics," Science, vol. 294, 2001

[22] M.Zamin Khan et al "Review of ASITIC (Analysis and Simulation of Inductors and Transformers for Integrated Circuits) Tool to Design Inductor on Chip" Vol 9 Issue 4, pg 196-201, 2012

[23] Shiraz Afzal, M.Zamin Ali Khan "A Proposed Silicon Optical Electronic Integrated Circuit with Monolithic Integration of LED, OPFET and Receiver Circuit". Vol 6 Issue 2, pg 42-46, 2012