Emerging Hardware Technologies and 3D System Integration for Ubiquitous Machine Intelligence

Haitong Li
Elmore Family School of Electrical and Computer Engineering, Purdue University
West Lafayette, IN, USA
haitongli@purdue.edu

2023 60th ACM/IEEE Design Automation Conference (DAC) | 979-8-3503-2348-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/DAC56929.2023.10247990

Abstract—**Next-generation semiconductor hardware technologies and system integration serve as the physical foundation in the pursuit of ubiquitous machine intelligence, with unprecedented requirements in energy efficiency, performance, cost effectiveness, and security. Here, we provide an overview of emerging technologies with an emphasis on 3D system integration, and discuss on cross-layer designs for memory-centric computing in the 3D era.**

Index Terms—**emerging device technologies, 3D integrated systems, memory-centric computing, artificial intelligence hardware.**

I. Introduction

Artificial intelligence is revolutionizing a broad spectrum of applications to address societal needs from computing to healthcare, with impact on global infrastructure across diverse sectors. Energy efficiency, carbon footprint, and privacy considerations [1] are rising as we continue to develop and deploy large and diverse artificial intelligence (AI) models across domains. The next-generation computing hardware must deliver drastic improvements and new functionalities. On the one hand, the historical density trends of semiconductor technologies, including logic, memory, and interconnect, show orders of magnitude of density up-scaling over nearly five decades [2], [3]. On the other hand, however, siloed research within traditional boundaries becomes insufficient. We usher in a new era where orchestrating a multitude of emerging device technologies with various 3D system integration schemes is needed to unlock domain-specific, cross-layer designs meeting ever-increasing application demands (Fig. 1).

II. Emerging Technology Landscape in 3D Era

A. System Integration as a Platform Technology

The system-level integration of dedicated component technologies provides a path for scaling and customization of any computing system. With the diminishing returns of two-dimensional down-scaling [4], [5], breaking the single-die limit and embracing the heterogeneity with the third dimension become crucial for meeting diverse design targets with improved cost effectiveness [6]–[8]. Manufacturing/integration of 2.5D/3D chiplets is now considered a critical 'platform technology' to support modular, flexible, and diverse designs in a full system with enhanced and enriched functionalities. The heterogeneous nature beyond conventional silicon is derived from several driving factors in addition to costs. First, enabling machine intelligence from edge to cloud poses disparate requirements for logic nodes, memory types, and/or specialty devices to be integrated in different ways unattainable on the conventional 2D chips. Second, the 'mix-and-match' feature allows designers to partition a system with different architectural options while exploiting the unique properties of underlying device technologies. For substrate interconnect fabrics, Si interposers or wafer-level fan-out layers are used in 2.5D/3D chiplets [7], [9]. For the inter-layer connections with the vertical stacking, process technologies including micro-bumps, through-silicon vias (TSV), hybrid bonding, and monolithic inter-layer vias (ILV), have been evolving with diverse characteristics (pitch size and density, reliability, parasitics, etc.) to deliver different power-performance-area-cost (PPAC) envelopes.

B. CMOS + X: Technology Enablers

The 2.5D/3D system integration platform makes various emerging device technologies promising not merely as 'replacements' but rather as 'enablers' for new architectures not readily available in conventional 2D designs. This is achieved through integration, aggregation and assembly of these evolving technologies along with silicon CMOS foundation to either augment or re-architect today's systems. Commonly referred to as 'CMOS + X', such technology integration can span several device categories, and can be manufactured in sequential/monolithic, heterogeneous, or a hybrid fashion, depending on the granularity requirements of dies/layers/circuits. For intelligent systems requiring end-to-end functionalities at the edge, silicon CMOS logic may be integrated with memories, sensors, actuators, RF/mm-wave, energy harvesting and power electronics. Emerging technologies in spintronics, photonics, or even quantum computing may even reside in a heterogeneous system with hybrid data representations.

Figure 1 illustrates the technology landscape with an example where a 3D stacked chip as part of a larger integrated/packaged system may integrate multiple logic and memory technologies in the back-end-of-line (BEOL) processes. The 3D layers may be fabricated in a BEOL process flow (under low-temperature processing) or bonded at die/wafer level if mixed-node silicon CMOS is needed. Dense on-chip memories [10], highly-scaled 2D-semiconductor FETs [11], and ultra-low-leakage oxide-semiconductor transistors [12] together serve

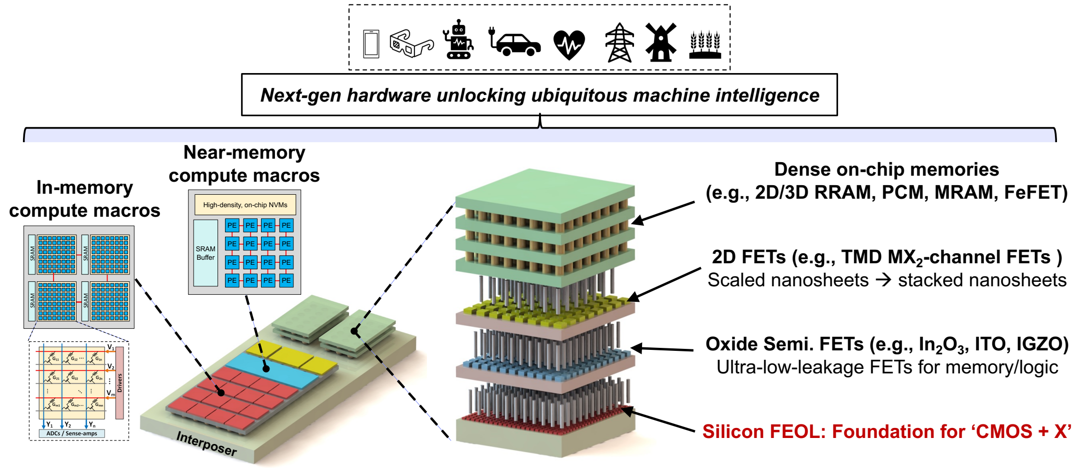
Next-gen hardware unlocking ubiquitous machine intelligence
In-memory compute macros
SRAM
V_1
V_2
V_m
G_{11}
G_{12}
G_{1n}
G_{21}
G_{22}
G_{2n}
G_{m1}
G_{m2}
G_{mn}
Drivers
Y_1
Y_2
Y_n
ADCs / Sense-amps
Near-memory compute macros
High-density, on-chip NVMs
SRAM Buffer
PE
Interposer
Dense on-chip memories (e.g., 2D/3D RRAM, PCM, MRAM, FeFET)
2D FETs (e.g., TMD MX_2-channel FETs)
Scaled nanosheets → stacked nanosheets
Oxide Semi. FETs (e.g., In_2O_3, ITO, IGZO)
Ultra-low-leakage FETs for memory/logic
Silicon FEOL: Foundation for 'CMOS + X'

Fig. 1. Overview of the emerging technology landscape in the 3D era towards. Energy-efficient, functionally-enriched computing systems that are closely designed with both application and technology characteristics in mind are key to unlocking future machine intelligence.

as 'domain-specific technologies' (DST) [3] offering a vast design space, where high-bandwidth, high-capacity, and energy-efficient architectures can be identified owing to the unique device characteristics and inter-connectivity that do not exist in conventional 2D Si-only systems. The BEOL integration and potentially stacking of nanosheets made of low-dimensional materials as well as nanoscale transistors made of wide-bandgap oxide semiconductors provide new scaling pathways, and may be utilized to construct new functional kernels beyond logic [13], [14]. 3D integration also exploits the benefits of dense on-chip memories beyond SRAM [6], with various genres such as resistive RAM (RRAM), phase change memory (PCM), magnetic RAM (MRAM), and ferroelectric FET (FeFET), which naturally lead to possibilities of architecting the 3D systems in the memory-centric fashion for data-intensive applications.

III. Memory-Centric Computing

The rich connectivity through 3D system integration opens up new opportunities for memory-centric computing: cross-layer co-designs can be realized by exposing and connecting the unique properties of emerging technologies at the device and circuit levels, to the diverse algorithm needs and characteristics. A recent example demonstrated on-chip one-shot learning leveraging cross-layer design and integration of RRAM and silicon CMOS [15]. As broadly illustrated in Fig. 1, heterogeneous integration allows partitioning workloads to combined near-memory/in-memory compute macros [16], [17]. With modular and flexible architectures, it is promising to further explore the synergistic mix-and-match of emerging computational paradigms, such as neuromorphic computing [18], probabilistic computing [19], and hyperdimensional computing [20], ultimately leading towards efficient neuro-symbolic AI systems.

Acknowledgment

We acknowledge fruitful discussions with Zhihong Chen, Ashraf Alam, Joerg Appenzeller, Peide Ye, Anand Raghunathan, and Kaushik Roy at Purdue University. We thank Luqi Zheng (Purdue) for creating the 3D illustrations in Fig. 1.

References

[1] E. Strubell *et al.*, "Energy and policy considerations for deep learning in NLP," *arXiv preprint arXiv:1906.02243*, 2019.
[2] H.-S. P. Wong *et al.*, "A density metric for semiconductor technology [point of view]," *Proceedings of the IEEE*, vol. 108, no. 4, 2020.
[3] K. Akarvardar *et al.*, "Technology prospects for data-intensive computing," *Proceedings of the IEEE*, vol. 111, no. 1, pp. 92–112, 2023.
[4] R. H. Dennard *et al.*, "Design of ion-implanted MOSFET's with very small physical dimensions," *JSSC*, 1974.
[5] T. N. Theis *et al.*, "The end of Moore's law: A new beginning for information technology," *Computing in Science & Engineering*, 2017.
[6] S. Salahuddin *et al.*, "The era of hyper-scaling in electronics," *Nature Electronics*, vol. 1, no. 8, pp. 442–450, 2018.
[7] C. Douglas *et al.*, "Foundry perspectives on 2.5 d/3d integration and roadmap," in *2021 IEEE IEDM*. IEEE, 2021, pp. 3–7.
[8] F. Sheikh *et al.*, "2.5 d and 3d heterogeneous integration: emerging applications," *IEEE Solid-State Circuits Magazine*, vol. 13, no. 4, 2021.
[9] R. Agarwal *et al.*, "3d packaging for heterogeneous integration," in *2022 ECTC*. IEEE, 2022, pp. 1103–1107.
[10] H.-S. P. Wong *et al.*, "Memory leads the way to better computing," *Nature nanotechnology*, vol. 10, no. 3, pp. 191–194, 2015.
[11] S. Das *et al.*, "Transistors based on two-dimensional materials for future integrated circuits," *Nature Electronics*, vol. 4, no. 11, 2021.
[12] M. Si *et al.*, "Scaled indium oxide transistors fabricated using atomic layer deposition," *Nature Electronics*, vol. 5, no. 3, pp. 164–170, 2022.
[13] R. Yang *et al.*, "Ternary content-addressable memory with MoS2 transistors for massively parallel data search," *Nature Electronics*, 2019.
[14] Y. Shi *et al.*, "Electronic synapses made of layered two-dimensional materials," *Nature Electronics*, vol. 1, no. 8, pp. 458–465, 2018.
[15] H. Li *et al.*, "SAPIENS: A 64-kb RRAM-based non-volatile associative memory for one-shot learning and inference at the edge," *IEEE Transactions on Electron Devices*, vol. 68, no. 12, pp. 6637–6643, 2021.
[16] A. Sebastian *et al.*, "Memory devices and applications for in-memory computing," *Nature nanotechnology*, vol. 15, no. 7, pp. 529–544, 2020.
[17] J. Gómez-Luna *et al.*, "Benchmarking memory-centric computing systems: Analysis of real processing-in-memory hardware," in *IGSC*, 2021.
[18] I. Chakraborty *et al.*, "Pathways to efficient neuromorphic computing with non-volatile memory technologies," *Applied Physics Reviews*, vol. 7, no. 2, 2020.
[19] S. Chowdhury *et al.*, "A full-stack view of probabilistic computing with p-bits: devices, architectures and algorithms," *IEEE JxCDC*, 2023.
[20] T. F. Wu *et al.*, "Hyperdimensional computing exploiting carbon nanotube FETs, resistive RAM, and their monolithic 3D integration," *IEEE JSSC*, vol. 53, no. 11, pp. 3183–3196, 2018.