CMOS+X Technologies for Neuro-Vector-Symbolic Computing

Luqi Zheng[1] and Haitong Li[1]
[1]*School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA*
Email: haitongli@purdue.edu / Phone: 765-496-0740

2024 Device Research Conference (DRC) | 979-8-3503-7373-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/DRC61706.2024.10605469

Introduction

Neuro-vector-symbolic (NVS) computing aims to combine neural models with vector-symbolic architectures (VSA), also known as hyperdimensional computing (HDC) built upon vector arithmetics in the high-dimensional space [1]. NVS computing merges human-like symbolic learning and reasoning with hierarchical, context-based neural representations, marking a key development in the third wave of artificial intelligence represented by Neuro-Symbolic AI [2]. With heterogeneous model architectures embedding diverse backbones and multiple data representations, NVS computing promise the next level of learning capability, robustness, and explainability. Deep neural networks (DNNs) [4] and spiking neural networks (SNNs) [5] in general, with specialized convolutional neural networks (CNNs) [3] and transformers [6], may form multi-modality neural backbones. VSA components [7][8][9] enhance cognitive reasoning, continuous learning, and domain adaptation. With NVS models handling data-intensive workloads, tailored hardware platforms are needed for sustained efficiency bridging the edge-cloud continuum. Targeting NVS-AI hardware, it is essential to have energy-efficient compute kernels optimized in a fine-grained fashion for various NVS components, as well as flexible architecture integration. These two critical needs can be met if we jointly exploit the 3D integration platforms and emerging device technologies with silicon CMOS foundation ("CMOS+X"), due to the vast design space exposed from material/device level to architecture level. We provide an overview of emerging device technologies which may enable system integration of hybrid neural and vector-symbolic models through NVS co-designs.

Device Technologies for Heterogeneous Nanokernels

A tailored architecture for hybrid NVS models may comprise several essential processing kernels, including digital multiply-accumulate (MAC), analog MAC, multiply-add-permute (MAP), activation and synaptic functions, distance measurement kernels (search), and efficient on-chip memories. Properly designed and realized with diverse device technologies, these "nanokernels" can be further integrated into 2D/2.5D/3D systems tailored for NVS architectures. Figure 1 illustrates, in a first-order fashion, an example mapping from selected device characteristics to sample NVS-friendly "CMOS+X" nanokernels. For neural models, near-/in-memory acceleration can leverage silicon CMOS and novel BEOL-compatible transistors such as ultra-thin oxide semiconductor FETs (OSFETs) and transition metal dichalcogenide FETs (TMD-FETs) to provide digital logic realization [10], and high-speed, low-leakage on-chip memories like embedded DRAM [12]. As a result, digital/analog MAC and activation kernels may be realized across FEOL and BEOL stacks [11]. Non-volatile memories such as resistive RAM (RRAM), phase change memory (PCM), magnetic RAM (MRAM), ferroelectric FET (FeFET), and electrochemical RAM (ECRAM) can be used to build analog MACs, activation, and synaptic kernels [4, 13, 14, 15], employing intrinsic physics and circuit-level properties. For VSA components, MAP and similarity measurement kernels are tailored specifically for high-dimensional vector operations, which inherently lead to memory-centric designs. Compact MAP kernels for HD computing may be realized with a mixed analog-digital blocks by integrating BEOL transistors and memory devices with excellent 3D scaling characteristics [9]. In-memory search operations for similarity measurement kernels can be efficiently realized with content-addressable memories (CAM), effectively utilizing the non-volatility and parallelism of BEOL NVM arrays [16]. Unique properties of materials and 3D scalability of devices may become the key pillars for future energy-area-efficient nanokernel designs for novel neuromorphic models.

The expansive design space offered by co-integration of CMOS+X technologies presents new opportunities for co-design of NVS-AI hardware with hybrid nature. Looking forward, cross-layer co-design activities and hardware prototyping (devices, circuits, and chiplets) must go hand in hand to truly materialize the promises and accelerate the lab-to-fab transition.

[1] D. Kleyko et al., ACM Comput. Surv., vol. 55, p. 1, (2021).
[2] A. D. Garcez et al., A.I. Reviwe, p. 1, (2023).
[3] Y. H. Chen et al., ISSCC, p. 262, (2016).
[4] W. Wan et al., Nature, vol. 608, p. 504, (2022).
[5] G. K. Chen et al., JSSC, vol. 54, p. 992, (2019).
[6] F. Tu et al., JSSC, vol. 58, p. 1798, (2023).
[7] S. Datta et al., JETCAS, vol. 9, p. 439, (2019).
[8] G. Karunaratne et al., Nat. Electron., vol. 3, p. 327, (2020).
[9] T. F. Wu et al., JSSC, vol. 53, p. 3183, (2018).
[10] W. Tang et al., IEDM, p. 483, (2022).
[11] Y. Wang et al., Nat. Commun., vol. 12, p. 3347, (2021).
[12] C. Chen et al., IEDM, p. 26.5.1, (2022).
[13] D. Satio et al., Symposium on VLSI Circuits, p. 1, (2021).
[14] Bianchi, S. et al., Nat. Commun., vol. 14, p. 1565, (2023).
[15] J. Cui et al., Nat. Electron., vol. 6, p. 292, (2023).
[16] T. Soliman et al., Nat. Commun., vol. 14, p. 6348, (2023).

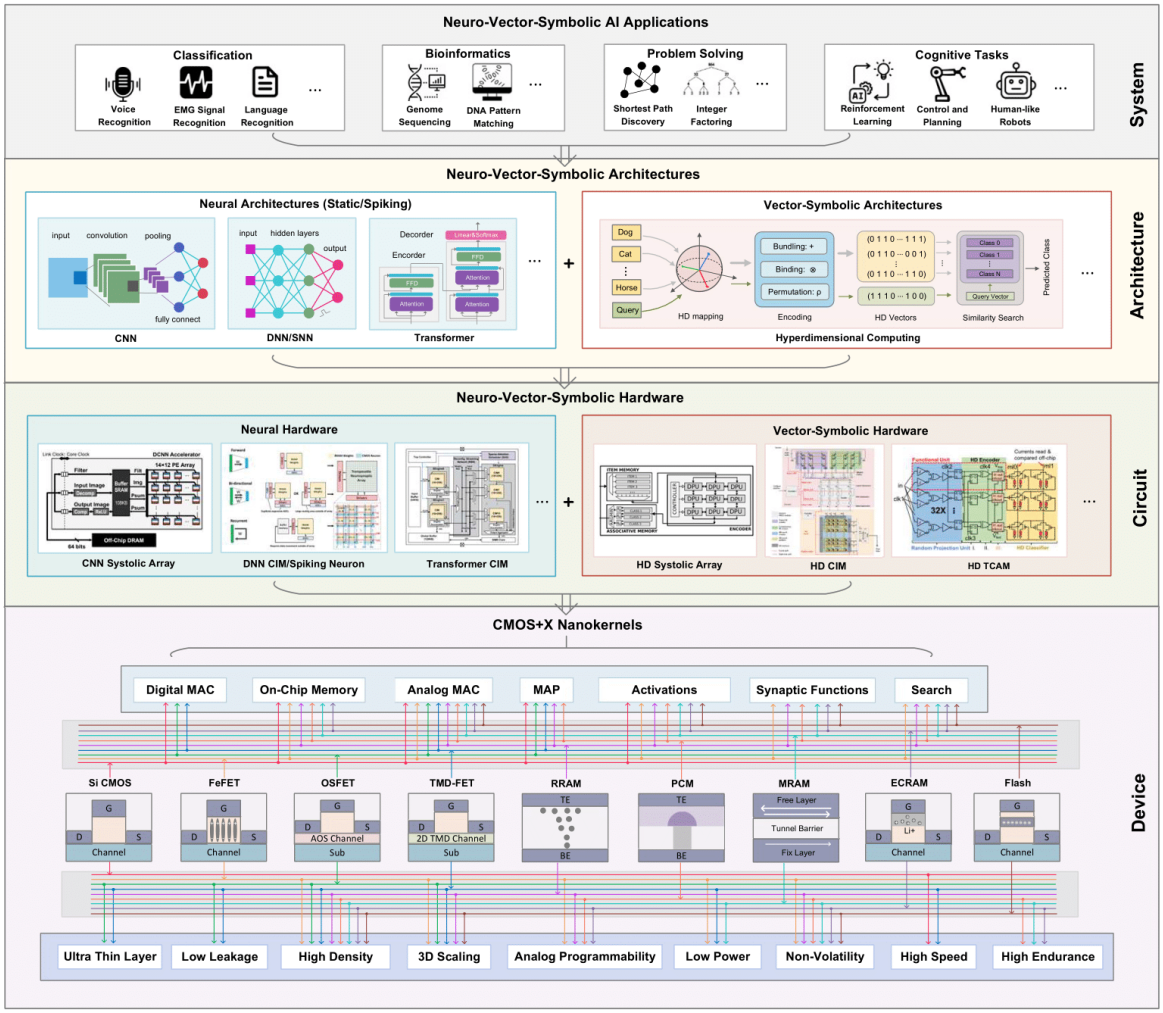

Fig. 1: **Co-design and co-integration of "CMOS+X" nanokernels serve as the foundation for neuro-vector-symbolic (NVS) computing**. With heterogeneous model architectures embedding diverse backbones and multiple data representations, NVS computing promise the next level of learning capability, robustness, and explainability. Tailored hardware platforms are needed for sustained efficiency in a cost-effective fashion bridging the edge-cloud continuum. Here, a first-order mapping from selected device characteristics to sample NVS-friendly "CMOS+X" nanokernels is shown: monolithic and heterogeneous integration of silicon CMOS and BEOL devices, along with dense off-chip storage, form a suite of efficient building blocks that require orchestration for tailored NVS architectures. The schematics in the circuit layer are adapted from [7][8][9].