Tutorial: The Synergy of Hyperdimensional and In-memory Computing

Paul R. Genssler[1], Simon Thomann[2, 3], and Hussam Amrouch[2, 3]
[1]University of Stuttgart, Stuttgart, Germany, [2]Chair of AI Processor Design, Technical University of Munich (TUM), Munich, Germany, [3]Munich Institute of Robotics and Machine Intelligence, TUM, Munich, Germany
Corresponding author: amrouch@tum.de

ABSTRACT

Breakthroughs in deep learning consistently drive innovation. However, DNNs tend to overwhelm conventional computing systems. Hyperdimensional Computing (HDC) is rapidly gaining prominence as a potent method for rapid learning from a relatively small amount of data. It also holds the promise of offering energy-efficient lightweight computation. This tutorial will provide a comprehensive overview of the major shortcomings of existing von Neumann architectures and the growing need for innovative designs that fundamentally reduce memory latency and energy consumption by enabling data processing within the memory itself. Additionally, the tutorial will delve into the immense potential of beyond-von Neumann architectures, which utilize both emerging beyond-CMOS devices like Ferroelectric Field-Effect Transistors (FeFET) and conventional CMOS devices.

CCS CONCEPTS

• **Computer systems organization → Special purpose systems**;
• **Hardware → Emerging technologies**; **Hardware reliability**;
• **Computing methodologies → Machine learning**.

KEYWORDS

Hyperdimensional Computing, In-memory Computing, Machine Learning, Reliability

ACM Reference Format:
Paul R. Genssler, Simon Thomann, and Hussam Amrouch. 2023. Tutorial: The Synergy of Hyperdimensional and In-memory Computing. In *2023 International Conference on Hardware/Software Codesign and System Synthesis (CODES/ISSS '23 Companion), September 17–22, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3607888.3608944

1 INTRODUCTION

Advances in deep learning regularly create remarkably innovations that enhance our day-to-day life. However, deep neural networks (DNNs) tend to overwhelm conventional computing systems due to substantial bottlenecks caused by data movement between processing units and memory. Consequently, the development of innovative and intelligent computing architectures becomes essential to augment or even replace the current von Neumann architecture, which has remained largely unchanged for decades. Moreover, deep learning requires substantial computing power to manage the massive number of multiply-accumulate (MAC) operations that must be performed simultaneously. As a result, there is an increasing demand for the development of innovative machine learning algorithms that emphasize learning from limited data and replacing costly MAC circuits with more straightforward and simpler logical operations. These advancements are crucial for realizing lightweight and efficient implementations.

Hyperdimensional Computing (HDC) is rapidly gaining prominence as a potent method for rapid learning from little data [1, 2, 3]. Its potential is particularly promising due to its exceptional ability to perform "reasoning" akin to human thinking, a capability often unattainable in traditional machine learning-based approaches. Furthermore, HDC excels at swift inference operations, fast learning, and excellent pattern matching. The lightweight operations enable rapid and efficient hardware implementations in contrast to DNNs [4, 5]. This development paves the way for on-chip learning, where data gathered at the edge can be utilized for training purposes. Additionally, HDC is inherently robust against noise in the data and the underlying computations. It represents data as large vectors with thousands of components. All components hold similar quantities of information, avoiding susceptible most significant bits. Many HDC operations are also per component, i.e., an incorrect calculation impacts only a single component.

Deep learning and HDC are data-intensive applications. In contrast, the von Neumann architecture was designed decades ago when most workloads were compute-intensive. The separation of memory and processing units was not a bottleneck. However, today's data-intensive applications expose the memory wall, which is a limiting factor in the efficiency of computing systems. Transferring data consumes orders of magnitude more energy than processing it. In-memory computing, on the other hand, focuses on performing computations directly within the memory, rather than transferring data back and forth between the processor and memory [6]. This approach significantly improves computational speed and efficiency, as it eliminates the bottleneck caused by data movement. However, one of the challenges associated with in-memory computing is its susceptibility to noise, either due to the underlying analog computations [7] or due to process variation from manufacturing [8]. Device endurance also becomes a concern if emerging non-volatile memories, such as Ferroelectric Field-Effect Transistors (FeFET), are employed [9]. This unreliability poses a

CODES/ISSS '23 Companion, September 17–22, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0289-1/23/09…$15.00
https://doi.org/10.1145/3607888.3608944

significant concern for traditional deep-learning applications that rely on accurate computation. Increasing the reliability of the hardware incurs large overheads that minimize the actual gains from in-memory computing and advanced technology nodes.

A powerful synergy exists between the robustness of HDC from the application level and the unreliable yet efficient computations from in-memory computing. By leveraging distributed and redundant representations, HDC can withstand memory failures and noise, effectively preserving the integrity of encoded information. Integrating HDC with in-memory architectures can mitigate the unreliability concerns associated with traditional in-memory computing, enhancing the overall robustness and fault tolerance of computational systems. This combination holds promise for applications where both speed and resilience are essential, enabling reliable *and* efficient processing of complex data sets.

2 OVERVIEW OF THE TUTORIAL

The tutorial bridges the gap between the latest innovations in the underlying technology and recent breakthroughs in computer architecture. It will demonstrate that HW/Sw co-design is key to achieve reliable and efficient hyperdimensional in-memory computing.

2.1 Beyond von Neumann Architectures

In the first part, the tutorial will offer a comprehensive overview of the major shortcomings of contemporary architectures and the ever-growing need for innovative designs that fundamentally reduce memory latency and energy consumption by enabling data processing within the memory itself. Additionally, it will delve into the immense potential of non-von Neumann architectures, utilizing both emerging beyond-CMOS devices like FeFET and conventional CMOS devices. The tutorial will elaborate on how novel HDC models can be effectively trained with limited and noisy data. Moreover, it will showcase the robustness of HDC models against errors, which is essential in fostering and realizing beyond-von Neumann architectures where errors are inevitable.

2.2 Hands-on Session

In the hands-on session of the tutorial, participants will be introduced to the concept of HDC and its applications in data analysis and classification. They will learn how to process datasets using hypervectors with various encoding methods and explore different similarity metrics. The tutorial will cover different techniques for generating hypervectors, such as the continuous indexing method, and demonstrate how to build models using them. Subsequently, participants will perform classification tasks employing these HDC models. Additionally, the tutorial will address how the impact of hardware errors on hypervectors and the resulting classification accuracy can be investigated. Participants will learn to model the influence of hardware errors on hypervectors and analyze the effects of these errors on classification.

By the end of the hands-on tutorial, participants will have a solid understanding of HDC and will be equipped with the skills to process datasets, perform classification, and analyze how hardware errors impact the training and inference of HDC algorithms.

3 PRESENTERS

Hussam Amrouch is Professor heading the Chair of AI Processor Design within the Technical University of Munich (TUM). He is also with the Munich Institute of Robotics and Machine Intelligence in Germany and also the head of the Semiconductor Test and Reliability (STAR) at the University of Stuttgart. Prior to that, he was a Research Group Leader at the Karlsruhe Institute of Technology. He currently serves as Editor at the Nature Scientific Reports Journal. He received his Ph.D. degree with the highest distinction from KIT in 2015. He has more than 210 publications in multidisciplinary research areas (including 86 journals), starting from semiconductor physics to circuit design all the way up to computer architectures. His research in HW security and reliability have been funded by the German Research Foundation, Advantest Corporation, and the U.S. Office of Naval Research (ONR).

Simon Thomann earned his degrees in Computer Science, Master in 2022 as well as Bachelor in 2019, at Karlsruhe Institute of Technology (KIT), Germany. He started his Ph.D. at the University of Stuttgart in 2022 and has been continuing it since 2023 at the Chair of AI Processor Design within the Technical University of Munich (TUM). His research interests range from device physics to the system level. His special interest lies in circuit design, emerging technologies, and cross-layer reliability modeling from device to circuit level.

Paul R. Genssler received the Dipl. Inf degree (M.Sc.) in computer science in 2017 at TU Dresden, Germany. In 2018 he started his PhD research at the Chair for Embedded Systems (CES) at Karlsruhe Institute of Technology, Germany. Since 2020 he continues his PhD with Prof. Amrouch at the Semiconductor Test and Reliability (STAR) chair within the Computer Science, Electrical Engineering Faculty at the University of Stuttgart. His research interests include emerging technologies, system architecture, and emerging brain-inspired methods for IC test and beyond.

ACKNOWLEDGMENTS

This research was supported in part by Advantest as part of the Graduate School "Intelligent Methods for Test and Reliability" (GS-IMTR) at the University of Stuttgart.

REFERENCES

[1] Paul R. Genssler et al. 2022. Brain-inspired computing for circuit reliability characterization. *IEEE Transactions on Computers*, 71, 12, (Feb. 2022), 3336–3348.

[2] Paul R. Genssler et al. 2021. Brain-inspired computing for wafer map defect pattern classification. In *IEEE International Test Conference (ITC'21)*. (Oct. 2021).

[3] Paul R. Genssler et al. 2023. Modeling and predicting transistor aging under workload dependency using machine learning. *IEEE Transactions on Circuits and Systems I: Regular Papers*.

[4] Simon Thomann et al. 2022. All-in-memory brain-inspired computing using FeFET synapses. *Frontiers in Electronics*, 3, (Feb. 2022).

[5] Simon Thomann et al. 2022. HW/SW co-design for reliable in-memory brain-inspired hyperdimensional computing. *IEEE Transactions on Computers*.

[6] Shan Deng et al. 2022. Compact ferroelectric programmable majority gate for compute-in-memory applications. In *2022 International Electron Devices Meeting (IEDM)*. IEEE, 36–7.

[7] Kai Ni et al. 2021. On the channel percolation in ferroelectric fet towards proper analog states engineering. In *2021 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 15–3.

[8] Simon Thomann et al. 2021. On the reliability of in-memory computing: impact of temperature on ferroelectric tcam. In *2021 IEEE 39th VLSI Test Symposium (VTS)*. IEEE, 1–6.

[9] Paul R. Genssler et al. 2022. On the reliability of fefet on-chip memory. *IEEE Transactions on Computers*, 71, 4, (Mar. 2022), 947–958.