Thesis Project Portfolio

Optimizing Convolutional Neural Networks on Processing-in-Memory Architectures: Implementation, Benchmarking, and Performance Analysis
(Technical Report)

Rethinking Performance: The Ethical and Technological Landscape of Accelerated Computing
(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Hugo Abbot

Spring, 2025

Department of Computer Science

Table of Contents

Executive Summary

Both my Capstone and STS research connect through their shared exploration of emerging technologies in computing, especially those that challenge the longstanding paradigms of processor performance and architecture. While my Capstone is a technical report on my work as part of a research lab studying Processing in Memory (PIM) technologies and their efficacy with convolutional neural networks (CNNs), my STS paper investigates the broader societal, ethical, and technological implications of accelerated computing hardware, including PIM along with other interesting developments throughout history. What unites both efforts is an undertaking to not only understand how these technologies function and provide benefit, but also how they evolve within and are shaped by societal, environmental, and economic contexts – creating an endless reciprocal relationship. This alignment between topics is driven by both academic and personal curiosity, studying the forces at play which guide innovation and the real-world impact of computing infrastructure.

This Capstone project investigates the performance of CNNs on emerging PIM architectures, focusing on detailed kernel-level analysis and end-to-end benchmarking. PIM represents a promising hardware paradigm that seeks to reduce the energy and performance bottlenecks caused by traditional memory-processor separation by embedding compute logic directly within memory modules. My work centered on the implementation and performance characterization of the ResNet-18 model using a custom high-level PIM simulator and benchmarking suite developed at The Laboratory for Computer Architecture at Virginia (LAVA Lab). This included manually constructing each layer using a C++ API, with particular attention to convolution, pooling, and activation layers, and benchmarking across multiple PIM configurations including bit-serial subarray-level, bit-parallel subarray-level, and bank-level

designs. The results showed that while PIM offers advantages in isolated compute performance, overall execution timing is often dominated by host-side processing and data movement overhead. Among the tested configurations, bit-parallel subarray-level PIM achieved the fastest in-memory compute times, though it still lagged behind traditional CPUs and GPUs in full-model execution. Toward the end of the project, we developed a new fully in-PIM convolution implementation, which removes the need for host interaction and showed up to a 10x speedup compared to the earlier partially offloaded version; however, the new convolution method has not yet been tested as part of the end-to-end ResNet-18 implementation, limiting our ability to accurately assess its overall impact. This work contributes to the broader vision of hybrid execution models by identifying which CNN kernels benefit most from PIM acceleration and lays the groundwork for future strategies that intelligently distribute workload across heterogeneous computing resources.

My STS research paper examines the societal consequences of the industry's shift away from traditional processing scaling in favor of technologies like GPUs, TPUs, and PIM. The paper employs Thomas P. Hughes' theory of Technological Momentum to contextualize how computing hardware evolves – first as malleable innovation, then as entrenched systems driven by market forces and historical decisions. Through two case studies – PIM and LK-99, a theorized room-temperature superconductor – I explore how ethical concerns such as environmental sustainability, technological inequalities, and accessibility emerge from the push for ever-greater efficiency. The study reveals how accelerated computing both responds to societal needs and reinforces existing power structures, highlighting challenges like E-waste, algorithmic bias, and uneven access to advanced tools. Ultimately, the paper argues that ethical

foresight must guide innovation to prevent the replication of historical inequities in emerging computing paradigms.

Working on both the Capstone project and STS research paper simultaneously allowed me to bridge technical insight with critical reflection in a way that neither experience could have offered on its own. My time in the research lab gave me first hand exposure to the obstacles faced in advancing a new computing architecture, including fabrication costs, design trade-offs, performance testing, and external collaboration – all of which grounding the theoretical discussions in my STS work. At the same time, the research paper pushed me to step back and consider the ethical, social, and global stakes of the very technologies we were exploring in the lab. Concepts like technological momentum, sustainability, and accessibility gained extra relevance when I saw how they played out in the day-to-day decisions made by us as researchers. The synergy between these projects helped me cultivate a more holistic engineering mindset, one that values not just innovation, but also responsibility, inclusion, and long-term impact. This integrative experience has reshaped how I think about the role of engineers in shaping the future and will help dictate how I engage with technology throughout my career.