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

Accelerated Computing Hardware: Examining Ethical Concerns and Sustainable Solutions

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Hugo Abbot

December 6, 2025

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor guidelines for Thesis-Related Assignments.

ADVISORS

Joshua Earle, Department of Engineering and Society

Kevin Skadron, Department of Computer Science

Introduction:

What would happen if the foundational principle that has guided computer architecture development for decades—Moore's Law—gradually became obsolete? Surprisingly, this potential decline has not sparked widespread panic, instead creating a new drive for alternate methods of research and performance increasing technology. Moore's Law, first articulated in 1965 by Gordon Moore, co-founder of Intel, posits that the number of transistors on integrated circuits doubles approximately every two years ("Moore's Law", 2023). While Moore himself never directly claimed that this trend would continue indefinitely, some, like NVIDIA's CEO Jensen Huang, have declared it over. Others, such as Intel's CEO Pat Gelsinger, maintain that Moore's Law is still "alive and well" (Leswing, 2022).

However, current barriers in processing development, such as physical limitations, rising energy costs, and other technical challenges, are pushing the field toward new channels of innovation. While these innovations aim to overcome technical constraints, they also introduce unforeseen consequences, including the potential to exacerbate social, economic, and structural inequalities. For instance, access to cutting-edge technology may increasingly be limited to wealthier nations or privileged groups, widening the gap between those with and without access to advanced computing power.

In this prospectus, I will explore the current ethical concerns surrounding the development of accelerated hardware in computing and how these issues are being addressed by the industry and stakeholders. These considerations include environmental sustainability, data privacy, resource inequality, and the sourcing of materials for production and maintenance. Understanding the ethical implications of accelerated hardware is crucial as this technology underpins many fields, from artificial intelligence to data centers and machine learning. Since

many of our future careers will build upon and utilize these systems, it is important to examine the various perspectives and implications of their use to help us become more socially responsible engineers.

Furthermore, I will examine the solutions being proposed to mitigate these challenges, including efforts to make hardware more energy-efficient, secure, and accessible. My analysis will be grounded in my ongoing research with the Laboratory for Computer Architecture at Virginia (LAVA Lab), which focuses on innovations in high-performance and energy-efficient computing systems.

Technical Project:

The growing demand for data processing in modern applications reveals performance bottlenecks in traditional DRAM-based architectures, particularly due to the narrow interface between memory and processors (i.e. CPUs or GPUs). These limitations result in significant data movement overhead, which reduces performance and energy efficiency (Jacob, 2010). Processing in Memory (PIM) minimizes data transfers by embedding computational capabilities into memory, leveraging its parallelism to enhance performance and energy efficiency, especially as traditional architectures face limitations due to the slowing of Moore's Law (Stone, 1970; Moore, 1998).

Recent innovations, such as standardized frameworks for evaluating different PIM architectures, allow for more direct comparisons and testing, pushing this technology towards the forefront of accelerated hardware development. However, various challenges persist, such as high data latency transfer costs and non-optimal instruction execution. For example, slow multiplication execution performance timings are present with bit-serial subarray-level PIM architecture types (Siddique, 2024).

My role as part of the LAVA Lab is multifaceted. I am responsible for maintaining and improving the benchmark and simulator tool for PIM simulation, as well as characterizing convolution neural networks (CNNs) on general-purpose PIM systems. The ladder will be my main focus and responsibility due to the rising interest and usage of artificial intelligence and machine learning models. Through this work, we aim to expand upon previous studies that have analyzed general machine learning workloads on memory-based computing systems (Gómez-Luna, 2023). PIM has shown significant speedup with various machine learning kernels (Blackford, 2002; Pearson, 1896), and with the increasing interest in CNNs and their applications in various industries, this area of research is both relevant and promising (Yamashita, 2018).

For this project, I will be responsible for the implementation, testing, and analysis of various kernels that make up these models. For example, ResNet-18, a model used in computer image recognition, consists of multiple layers, such as average pooling, max pooling, ReLU activation, among others ("ResNet18"). Each of these layers will be implemented using PIM devices with our high-level C++ simulator, and performance comparisons will be made to CPU and GPU baselines using highly optimized external libraries, such as CUDA, Boost, OpenBLAS. After determining which portions PIM can perform better, we plan to create a "Frankenstein-like" implementation where different kernels of a single model are handled by different processors, such as ReLU performed by the GPU while max pooling is run with PIM.

Following this submission, I will also be responsible for the design and implementation of a channel-level near-memory processing unit, similar to Intel's Data Streaming Accelerator ("Intel Data Streaming Accelerator", 2024). Previous studies have demonstrated the device's improved performance and efficiency across a wide range of applications (Kuper, 2024), and it will be incorporated into our benchmark and simulator suite to further evaluate the efficacy of

differing PIM types.

STS Project:

For this project, my research focuses on accelerated hardware in computing, including high-performance processors like CPUs, GPUs, and TPUs. These devices are designed to enhance computational performance and capability for applications such as artificial intelligence, data centers, and machine learning. The main question driving my exploration is: *What are the ethical concerns associated with the development of accelerated hardware, and how are these concerns being addressed by industry leaders*? The high-performance computing industry was valued at $50 billion in 2023 and is projected to grow to $110 billion by 2032, according to recent studies, underscoring the interest and investment in new innovation and infrastructure ("High Performance Computing Market Size & Trends [2030]", 2024). Emerging paradigms like accelerated hardware are often considered high risk for research due to their immaturity and ambiguous foundational principles, making it crucial to adopt a holistic strategy for continued development and evaluation (Buyya, 2021). As with past technologies, new innovations solve field-relevant problems but also introduce unintended consequences. By addressing these ethical challenges, we can better ensure that technological advancements benefit society as a whole, rather than reinforcing existing social, economic, and class disparities.

Many key groups and individuals are involved in the evolving landscape of computing advancements and are necessary to identify due to their influence and impact on the industry. Major tech companies, such as NVIDIA, AMD, and Intel, are some of the forefront stakeholders in the development of these new technologies with government agencies, such as the Defense Advanced Research Projects Agency (DARPA), Department of Defense (DOD), and the Armed Forces branches, interested for use in national security and technological colonialism (Sun,

2022). Additionally, Google and AWS, among other data center and cloud computing companies, use this technology to run large-scale operations, driving concerns for energy consumption. Policymakers possess the ability to shape regulations potentially limiting both positive and negative impacts, but must balance national and public interests, which has been explored previously in regards to automation as a whole (Yu, 2024). Finally, the awareness of these issues, much of what I'm trying to do in this work, needs to be addressed at the institutional level, starting with greater expanse of the Science, Technology, and Society (STS) field nationwide. Much of this exploration aims to find the many ways computer development impacts society; however, many groups, like certain minority groups or small-scale end users, may be left out from this discussion due to the wide-ranging effects and continuous development of computational systems nowaday; however, their accidental exclusion must be dealt with as best as possible as they can play a massive role in greater social acceptance of technology (Nawar, 2022).

Methodology and Frameworks:

Due to the nature of the research topic, much of my research will involve examining the historic elements of this evolving landscape, tracking the progress of advancements and their philosophical reasoning. By reviewing previous discussions on this topic through research studies, statistical analysis, and individual testimony, I will be able to draw stronger connections between past and present ethical dilemmas in computing, facilitating better problem-solving and preparation for future challenges. Another key method will be analyzing trends in public policy, particularly how policy shapes and encourages innovation, especially in response to changing industry standards, such as the slowing of Moore's Law. Additionally, I will analyze how these trends have influenced the ethical implications of accelerated hardware development, such as its

impact on society and the environment. Engaging with non-stakeholder discussions could also be important for understanding how new technology might inadvertently affect areas of life that were previously overlooked.

For my analysis, I will use the Evolution of Large Technical Systems framework, often referred to as Technological Momentum, which provides a deeper examination of the evolution of accelerated hardware and how it gains momentum at key points in history, ultimately becoming entrenched in society. This framework will allow me to explore how modern computing infrastructures have become dependent on these advancing technologies, creating ethical challenges that are increasingly difficult to address or reverse. Additionally, previous studies have applied this framework to examine the impact of innovative technologies, specifically in American society, providing a broader scope for analyzing the communities and groups affected by the rise of accelerated hardware (Song, 2024). Furthermore, I will use utilitarian ethics in my analysis which can give interesting insights into whether the overall benefits of innovation justify the potential risks (Green, 2016).

Throughout the upcoming months, I will conduct a thorough literature review of various academic papers, reports, and news articles that detail the issues and ethical concerns associated with accelerated hardware. Specifically, I will focus on areas such as environmental impact, including energy demands, privacy risks, and decreasing accessibility, as well as the uneven distribution of benefits to the public due to high costs. By January 2025, I will analyze media coverage and public perception and identify policy networks that govern the ethical regulation of this technology. By the end of February 2025, I aim to have completed most, if not all, of my research, including an analysis of at least one case study of ethical challenges caused by an innovative product and the solutions taken to address them. At that point, I will begin drafting

the final paper and submit it in early spring. I hope my work leads to better future design and considerations in new technologies, opening our eyes as engineers and citizens to the unforeseen consequences in the pursuit of computational advancement.

Key Texts

Buyya, R., Gill, S. S., Narayana, S. S., Bahsoon, R., & Murugesan, S. (2021). *A Strategy for Advancing Research and Impact in New Computing Paradigms*. ArXiv.org. https://arxiv.org/abs/2104.04070

This study discusses the challenges and opportunities in advancing research within new computing paradigms, focusing on high-risk areas that require thoughtful consideration and approaches for development. Using previous experience and knowledge from within the industry, the authors plan to use various methods, such as community outreach efforts and roadmap/simulator models to improve the impact of their research and adoption within society.

This work shows a case which is relevant in the context of my problem, showing an approach of minimizing potential problems with immature technologies. Additionally, it highlights the importance of continued and long-term observation in order to adequately assess societal impacts of the innovation, which are important topics I'd like to touch on within my work.

Woods, A. (2021). *The Death of Moore's Law: What it means and what might fill the gap going forward. MIT CSAIL Alliances.* https://cap.csail.mit.edu/death-moores-law-what-it-means-and-what-might-fill-gap-going-forward

This article explores the ending of Moore's Law, which has guided computer innovation for decades. Due to previously expected performance gains from semiconductor advancements ending, the authors argue that performance gains must come from higher up in the computing stack, such as in software, algorithms, and architecture.

The content from this discussion is highly relevant, providing both background information for my history aspect, as well as providing alternative technologies I can analyze and potentially put in my research paper. Additionally, ethical concerns are brought up, such as the need for more affordable products, greater access to education in this field, and the financial incentives for large corporate giants.

Song, D. (2024). *The Impact of Newly Innovative Technologies in American Society*. University of Virginia.

Song argues that the implementation of AI will have significant transformative effects on American society, similar to past technological advancements like the assembly line and personal computer.

This paper connects directly with the topic area I want to explore, specifically covering social and economic shifts triggered by new technology. The use of technological determinism is an interesting aspect and could be a useful aspect to examine when looking at large scale societal impact from new computer hardware innovation.

Yu, J. (2024). *The effects of automation on employment and government policy in the United States*. University of Virginia.

Yu explores the impact of automation technology and its impact on employment, specifically looking at the role of American policymakers in addressing it. The paper drives the argument and support of government intervention to mitigate job loss caused by the advancement, focusing on the manufacturing industry as an example.

The work from this provides valuable context and insight in examining how new hardware design and innovation can disrupt existing social and economic systems. This

consideration is one I plan hitting on during my research work, and appreciate the insight such as shifts in unemployment and wage stagnation.

References

Bannon, J. (2023). *How have developments in network technologies changed the relationship between users and e-commerce companies?* University of Virginia.

Blackford, I. S., Petitet, A., Pozo, R., Remington, K., Whaley, R. C., Demmel, J., Dongarra, J., Duff, I., Hammarling, S., Henry, G., & et al. (2002). An updated set of basic linear algebra subprograms (BLAS). *ACM Transactions on Mathematical Software (TOMS)*, *Volume 28*, pp. 135-151.

Buyya, R., Gill, S. S., Narayana, S. S., Bahsoon, R., & Murugesan, S. (2021). *A Strategy for Advancing Research and Impact in New Computing Paradigms*. ArXiv.org. https://arxiv.org/abs/2104.04070

Gómez-Luna, J., Guo, Y., Brocard, S., Legriel, J., Cimadomo, R., & Oliveira, G. F. (2023). Evaluating Machine Learning Workloads on Memory-Centric Computing Systems. *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pp. 35-49. https://ieeexplore.ieee.org/document/10158216

Green, B. (2016, December 16). *Exploring Ethics, Social Robots, and Artificial Intelligence*. Markkula Center for Applied Ethics at Santa Clara University. https://www.scu.edu/ethics/focus-areas/technology-ethics/resources/social-robots-ai-and-ethics/

High Performance Computing Market Size & Trends [2030]. (2024, September 30). Fortune Business Insights. https://www.fortunebusinessinsights.com/industry-reports/high-performance-computing-hpc-and-high-performance-data-analytics-hpda-market-100636

Intel Data Streaming Accelerator (Intel DSA). (2024). Intel.

https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/data-streaming-accelerator.html

Jacob, B., Wang, D., & Ng, S. (2010). *Memory systems: cache, DRAM, disk*.

Kuper, R., Jeong, I., Yuan, Y., Hu, J., Wang, R., Ranganathan, N., & Kim, N. S. (2024). A Quantitative Analysis and Guidelines of Data Streaming Accelerator in Modern Intel Xeon Scalable Processors. *In Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, *Volume 2*, pp. 37-54. https://arxiv.org/abs/2305.02480v5

Leswing, K. (2022, September 27). *Intel says Moore's Law is still alive and well. Nvidia says it's ended.* CNBC. https://www.cnbc.com/2022/09/27/intel-says-moores-law-is-still-alive-nvidia-says-its-ended.html

Moore, G. E. (1998). Cramming more components onto integrated circuits. *Proceedings of the IEEE*, *Volume 86*, pp. 82-85.

Moore's Law. (2023, September 18). Intel. https://www.intel.com/content/www/us/en/newsroom/resources/moores-law.html#gs.g68wmj

Nawar, W. (2022). *Role of Social Context in Technological Advancement*. University of Virginia.

Pearson, K. (1896). Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *no. 187*, pp. 253-318.

ResNet18 & ResNet50 in Computer Vision. (n.d.). Product Teacher.

https://www.productteacher.com/quick-product-tips/resnet18-and-resnet50

Siddique, F. A., Guo, D., Fan, Z., Gholamrezaei, M., Baradaran, M., Ahmed, A., Abbot, H., Durrer, K., Nandagopal, K., Ermovick, E., Kiyawat, K., Gul, B., Mughrabi, A., Venkat, A., & Skadron, K. (2024). Architectural Modeling and Benchmarking for Digital DRAM PIM. *IEEE International Symposium on Workload Characterization (IISWC)*. http://www.cs.virginia.edu/~skadron/Papers/PIMbench_PIMeval_iiswc2024.pdf

Song, D. (2024). *The Impact of Newly Innovative Technologies in American Society*. University of Virginia.

Stone, H. S. (1970). A logic-in-memory computer,. *IEEE Transactions on Computers (TC)*, *Volume 100*, pp. 73-78.

Sun, A. (2022). *The Geopolitics of Sociotechnical Systems: America's Digital Colonialism and China's Isolated Internet* . https://libraetd.lib.virginia.edu/downloads/1c18dg751?filename=Sun_Anthony_STS_Research_Paper.pdf

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights Imaging*, *9*, pp. 611-629. https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9