Department of Electrical Engineering
香港城市大學
City University of Hong Kong

McMaster University

CityU Architecture Lab for Arithmetic and Security (CALAS) Seminar Series

When Commodity Computing Tap Out, Custom Hardware Saves the Day: Unlocking High-Performance, Power-Efficient Solutions for Deep Learning and Neurotechnology

Dr. Ameer M. S. Abdelhadi, Assistant Professor
McMaster University, Hamilton, Canada

Abstract:

Commodity computing often fails to meet the stringent requirements of emerging compute- and memory-intensive applications, particularly those constrained by latency, form-factor, and energy—especially when deployed on portable and wearable devices. The success of such systems hinges upon real-time processing, energy efficiency, and adaptability to user needs. Existing solutions are often post-hoc, resulting in limited portability and energy efficiency. For such systems to become practical, highly optimized algorithms, innovative computing paradigms, and customized hardware are essential to enable portable, scalable, adaptable, real-time, and power-efficient processing. Two decades ago, the semiconductor technology faced the “power wall” due to the demise of Dennard scaling, which marked the end of the golden era where increasing frequency improved performance while reducing voltage maintained power consumption. Instead, massively parallel architectures emerged, driven by the continuity of Moore's Law, which predicted the availability of densely packed, cheaper transistors. A prime example of this paradigm shift is the field of deep learning. Although the foundational concepts of deep learning have been known for decades, their practical application was impeded by the limited capabilities of commodity hardware. The introduction of general-purpose GPUs, with their massive spatially parallel processing capabilities, revolutionized deep learning by enabling the efficient training and inference of complex models. However, as deep learning algorithms become more sophisticated and demanding, and as workloads continue to expand, even the most advanced GPUs are reaching their limits. This necessitates the adoption of custom accelerators specifically designed for deep learning tasks, particularly for portable applications, where GPUs fail in energy efficiency. This talk will delve into two critical domains where custom acceleration is indispensable: deep learning and neurotechnology. We will overview these fields and their potential applications and focus on the integral role that custom computer architecture has to play for these areas to flourish. By examining case studies, we will explore how custom-tailored hardware solutions have unlocked unprecedented performance and energy efficiency, advancing the capabilities of both fields. In neurotechnology, after taking a mile-high view of a brain-machine interface system, we will then highlight the unique challenges and solutions that have emerged in creating hardware to meet the stringent requirements of neural data processing. In deep learning, we will discuss our current hardware-accelerated deep learning techniques for reducing computation, data traffic, and memory footprint. We conclude with a review of future research directions for custom-tailored computer architecture and its applications.

Biography:

Ameer M. S. Abdelhadi is an assistant professor of Computer Engineering in the Department of Electrical and Computer Engineering at McMaster University. He obtained his PhD in Computer Engineering from the University of British Columbia in 2016. Prior to joining McMaster, Dr. Abdelhadi held various academic positions as a research fellow and lecturer at the University of Toronto, Imperial College London, and Simon Fraser University. Before pursuing his graduate studies, he held multiple design and research positions in the semiconductor industry. Dr. Abdelhadi’s research interests span multiple areas, including application-specific custom-tailored computer architecture and hardware acceleration, hardware-efficient deep learning, neurotechnology, reconfigurable computing, and asynchronous circuits.

1 Nov 2024 (Fri); 9:00am - 11:00am HKT; P1402; https://cityu.zoom.us/j/96742093029