Let's Go Sparse: Towards Sparse-Aware Computing Design

Christina Giannoula

Postdoctoral Researcher

University of Toronto

Tuesday, June 18

2:30 PM • MSEE 239

Zoom: https://purdue-edu.zoom.us/j/91038736134 ~ Meeting ID: 910 3873 6134

Abstract

Sparse computational kernels comprise an increasingly important workload domain for many fields, including bioinformatics, chemistry, web graph analysis, graph processing and are becoming ubiquitous in deep learning applications. Sparse workloads exhibit unique characteristics and very random data access patterns, which result in critical performance challenges in modern computing systems. In this talk, I will describe the challenges introduced in sparse kernel executions and I will open up several opportunities in the software, system, and hardware level to tackle sparsity. First, I will present an algorithmic engineering approach to improve performance of sparse convolution, a widely used sparse machine learning operator, in server-class GPUs. Second, I will present an adaptive hardware mechanism to mitigate data movement costs of low-locality sparse applications running on disaggregated systems. Finally, I will briefly describe a few design choices for efficient execution of sparse computational kernels in Processing-In-Memory systems. I will conclude by describing and advocating a principled approach to tackle sparsity that can enable us to more efficiently execute such important computational kernels.

Bio

Christina Giannoula is a Postdoctoral Researcher at the University of Toronto working with Prof. Gennady Pekhimenko, Prof. Andreas Moshovos and Prof. Nandita Vijaykumar and their research groups. She is also an affiliated senior researcher at the SAFARI research group, ETH Zurich, and is working with Prof. Onur Mutlu. Her current research interests lie in the intersection of computer architecture, computer systems and high-performance computing. Specifically, her research specializes on improving the performance and efficiency of emerging applications, with a focus on machine learning and sparse workloads, in modern computing paradigms, such as AI-specific GPU and Processing-In-Memory architectures, via software, system and hardware co-design. Christina has been honored with the 2023 Vector Institute PostDoc Research Grant, the 2023 Iakovos Giurunlian PhD Thesis Award and the 2022 Foundation for Education and European Culture PhD award. She also received a PhD Fellowship from 2017 to 2020, supported by the General Secretariat for Research and Technology (GSRT) and the Hellenic Foundation for Research and Innovation (HFRI). She earned her Ph.D. in October 2022 from School of Electrical and Computer Engineering (ECE) at the National Technical University of Athens (NTUA), advised by Prof. Georgios Goumas, Prof. Nectarios Koziris and Prof. Onur Mutlu.

PURDUE UNIVERSITY | Elmore Family School of Electrical and Computer Engineering