Geraldo Francisco de Oliveira Junior

Research Assistant at ETH Zurich

geraldod@safari.ethz.ch
Up-to-date version of CV is available at https://geraldofojunior.github.io/geraldofojunior.github.io

I am a Ph.D. candidate in the Safari Research Group at ETH Zürich, working with Prof. Onur Mutlu. My current broader research interests are in computer architecture, systems, and applications with a special focus on memory technologies and emerging applications. In particular, my Ph.D. research focuses on taking advantage of new memory technologies to accelerate distinct classes of applications. I have published several works on this topic in major venues such as ASPLOS, MICRO, and ISCA.

Education

2024 (expected)
PhD in Information Tech. and Electrical Engineering, ETH Zürich
2017
MSc in Computer Science, Federal University of Rio Grande do Sul
2015
BSc in Computer Science, Federal University of Vicosa

First-Author Publications

Heterogeneous Data-Centric Architectures for Modern Data-Intensive Applications: Case Studies in Machine Learning and Databases in ISVLSI 2022

Full Reference: Geraldo F. Oliveira, Amirali Boroumand, Saugata Ghose, Juan J. Gómez-Luna and Onur Mutlu, **"Heterogeneous Data-Centric Architectures for Modern Data-Intensive Applications: Case Studies in Machine Learning and Databases,"** IEEE Computer Society Annual Symposium on VLSI (ISLVSI), Nicosia, Cyprus, 2022.

We showcase the benefits of co-designing algorithms and hardware in a way that efficiently takes advantage of the PIM paradigm for two modern data-intensive applications: (1) machine learning inference models for edge devices and (2) hybrid transactional/analytical processing databases for cloud systems. We follow a two-step approach in our system design. In the first step, we extensively analyze the computation and memory access patterns of each application to gain insights into its hardware/software requirements and major sources of performance and energy bottlenecks in processor-centric systems. In the second step, we leverage the insights from the first step to co-design algorithms and hardware accelerators to enable high-performance and energy-efficient data-centric architectures for each application.

Full Paper: arXiv:
memory dram processing-in-memory machine learning databases

Methodologies, Workloads, and Tools for Processing-in-Memory: Enabling the Adoption of Data-Centric Architectures in ISVLSI 2022

Full Reference: Geraldo F. Oliveira, Juan J. Gómez-Luna, Saugata Ghose and Onur Mutlu, **"Methodologies, Workloads, and Tools for Processing-in-Memory: Enabling the Adoption of Data-Centric Architectures,"** IEEE Computer Society Annual Symposium on VLSI (ISLVSI), Nicosia, Cyprus, 2022.

Our goal in this work is to provide tools and system support for PnM and PuM architectures, aiming to ease the adoption of PIM in current and future systems. With this goal in mind, we address two limitations of prior works related to (i) identifying and characterizing workloads suitable for PnM offloading and (ii) enabling complex operations in PuM architectures. First, we develop a methodology, called DAMOV, that identifies sources of data movement bottlenecks in applications and associates such bottlenecks with PIM suitability. Second, we propose

an end-to-end framework, called SIMDRAM, that enables the implementation of complex inDRAM operations transparently to the programmer.

Full Paper: arXiv:
memory dram processing-in-memory benchmarking frameworks

Accelerating Neural Network Inference With Processing-in-DRAM: From the Edge to the Cloud in ISVLSI 2022

Full Reference: Geraldo F. Oliveira, Juan J. Gómez-Luna, Saugata Ghose and Onur Mutlu, **"Accelerating Neural Network Inference With Processing-in-DRAM: From the Edge to the Cloud,"** IEEE Micro, vol. 42, no. 6, pp. 25-38, 1 Nov.-Dec. 2022.

Neural networks (NNs) are growing in importance and complexity. An NN's performance (and energy efficiency) can be bound either by computation or memory resources. The processing-in-memory (PIM) paradigm, where computation is placed near or within memory arrays, is a viable solution to accelerate memory-bound NNs. However, PIM architectures vary in form, where different PIM approaches lead to different tradeoffs. Our goal is to analyze, discuss, and contrast dynamic random-access memory (DRAM)-based PIM architectures for NN performance and energy efficiency. To do so, we analyze three state-of-the-art PIM architectures: 1) UPMEM, which integrates processors and DRAM arrays into a single 2-D chip, 2) Mensa, a 3-D-stacking-based PIM architecture tailored for edge devices, and 3) SIMDRAM, which uses the analog principles of DRAM to execute bit-serial operations. Our analysis reveals that PIM greatly benefits memory-bound NNs: 1) UPMEM provides 23x the performance of a high-end graphics processing unit (GPU) when the GPU requires memory oversubscription for a general matrix-vector multiplication kernel, 2) Mensa improves energy efficiency and throughput by 3.0x and 3.1x over the baseline Edge tensor processing unit for 24 Google edge NN models, and 3) SIMDRAM outperforms a central processing unit/graphics processing unit by 16.7x/1.4x for three binary NNs. We conclude that the ideal PIM architecture for NN models depends on a model's distinct attributes, due to the inherent architectural design choices.

Full Paper: arXiv:
memory dram processing-in-memory neural networks inference

SIMDRAM: An End-to-End Framework for Bit-Serial SIMD Computing in DRAM in ASPLOS 2021

Full Reference: Nastaran Hajinazar, Geraldo F. Oliveira, Joao Dinis Ferreira, Nika Mansouri Ghiasi, Minesh Patel, Mohammed Alser, Saugata Ghose, Juan Gomez-Luna, and Onur Mutlu, **"SIMDRAM: An End-to-End Framework for Bit-Serial SIMD Computing in DRAM"** Proceedings of the 26th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), Virtual, March-April 2021.

Processing-using-DRAM has been proposed for a limited set of basic operations (i.e., logic operations, addition). However, in order to enable full adoption of processing-using-DRAM, it is necessary to provide support for more complex operations. In this paper, we propose SIMDRAM, a flexible general-purpose processing-usingDRAM framework that (1) enables the efficient implementation of complex operations, and (2) provides a flexible mechanism to support the implementation of arbitrary user-defined operations. We design the hardware and ISA support for SIMDRAM framework to (1) address key system integration challenges, and (2) allow programmers to employ new SIMDRAM operations without hardware changes.

Full Paper: arXiv:
bulk bitwise operations processing-in-memory dram memory

DAMOV: A New Methodology and Benchmark Suite for Evaluating Data Movement Bottlenecks in IEEE Access 2021

Full Reference: Geraldo F. Oliveira , Juan Gomez-Luna, Lois Orosa, Saugata Ghose, Nandita Vijaykumar, Ivan fernandez, Mohammad Sadrosadati and Onur Mutlu, **"DAMOV: A New Methodology and Benchmark Suite for Evaluating Data Movement Bottlenecks,"** IEEE Access, vol. 9, pp. 134457-134502, 2021.

Our goal is to methodically identify potential sources of data movement over a broad set of applications and to comprehensively compare traditional compute-centric data movement mitigation techniques (e.g., cach ing and prefetch ing) to more memory-centric techniques (e.g., NDP), thereby developing a rigorous understanding of the best techniques to mitigate each source of data movement. With this goal in mind, we perform the first large-scale characterization of a wide variety of applications, across a wide range of application domains, to

identify fundamental program properties that lead to data movement to/from main memory. We develop the first systematic methodology to classify applications based on the sources contributing to data movement bottlenecks.

Full Paper: arXiv: Short Talk (21 mins): Long Talk (2 hrs 40 mins):

benchmarking | data movement | memory systems | dram | near-data processing | workload characterization

Extending Memory Capacity in Consumer Devices with Emerging Non-Volatile Memory: An Experimental Study in arXiv 2021

Full Reference: Geraldo F. Oliveira , Saugata Ghose, Juan Gómez-Luna, Amirali Boroumand, Alexis Savery, Sonny Rao, Salman Qazi, Gwendal Grignou, Rahul Thakur, Eric Shiu and Onur Mutlu **"Extending Memory Capacity in Consumer Devices with Emerging Non-Volatile Memory: An Experimental Study","** arXiv:2111.02325 [cs.AR], 2021.

In this work, we provide the first analysis of the impact of extending the main memory space of consumer devices using off-the-shelf NVMs. We extensively examine system performance and energy consumption when the NVM device is used as swap space for DRAM main memory to effectively extend the main memory capacity. For our analyses, we equip real web-based Chromebook computers with the Intel Optane SSD, which is a state-of-the-art low-latency NVM-based SSD device. We compare the performance and energy consumption of interactive workloads running on our Chromebook with NVM-based swap space, where the Intel Optane SSD capacity is used as swap space to extend main memory capacity, against two state-of-the-art systems: (i) a baseline system with double the amount of DRAM than the system with the NVM-based swap space; and (ii) a system where the Intel Optane SSD is naively replaced with a state-of-the-art (yet slower) off-the-shelf NAND-flash-based SSD, which we use as a swap space of equivalent size as the NVM-based swap space.

Full Paper:

consumer devices | emerging memory | web browsing | dram | optane | chromebook

Employing Classification-based Algorithms for General-Purpose Approximate Computing in DAC 2018

Full Reference: Geraldo F. Oliveira, Larissa Rozales Gonçalves, Marcelo Brandalero, Antonio Carlos S. Beck and Luigi Carro, **"Employing Classification-based Algorithms for General-Purpose Approximate Computing,"** in Proceedings of the 55th Annual Design Automation Conference (DAC), San Francisco, CA, USA, 2018.

Approximate computing has recently reemerged as a design solution for additional performance and energy improvements at the cost of output quality. In this paper, we propose using a tree-based classification algorithm as an approximation tool for general-purpose applications. We show that, without any hardware support, completely implemented in software, our approach can improve performance by up to 4x (1.95x on average) and reduce EDP by up to 19x (4.04x on average) when compared to precise executions. Besides that, in some cases, our software-based mechanism can even outperform traditional hardware-based Neural Network's state-of-the-art designs.

Full Paper:

approximate computing | classification | decision trees

NIM: An HMC-Based Machine for Neuron Computation in ARC 2017

Full Reference: Geraldo F. Oliveira, Paulo C. Santos, Marco A. Z. Alves and Luigi Carro, **"NIM: An HMC-Based Machine for Neuron Computation","** in Proceedings of the 13th International Symposium of pplied Reconfigurable Computing (ARC), Delft, The Netherlands, 2017.

Neuron Network simulation has arrived as a methodology to help one solve computational problems by mirroring behavior. However, to achieve consistent simulation results, large sets of workloads need to be evaluated. In this work, we present a neural in-memory simulator capable of executing deep learning applications inside 3D-stacked memories. With the reduction of data movement and by including a simple accelerator layer near to memory, our system was able to overperform traditional multi-core devices, while reducing overall system energy consumption.

Full Paper:

processing-in-memory | near-data processing | neuron simulation | neural networks

A Generic Processing in Memory Cycle Accurate Simulator Under Hybrid Memory Cube Architecture in SAMOS 2017

Full Reference: Geraldo F. Oliveira, Paulo C. Santos, Marco A. Z. Alves and Luigi Carro, **"A Generic Processing in Memory Cycle Accurate Simulator Under Hybrid Memory Cube Architecture","** in Proceedings of the International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS), Pythagorion, Greece, 2017.

In this paper, we show the development of a precise, modular and parametrized PIM simulation environment. Our simulator has been developed using the SystemC allowing native parallel simulation. We have implemented the latest HMC technical specifications, including all HMC instructions. The primary contribution of our work lies on developing a user-friendly interface to allow easy PIM architectures exploitation. To evaluate our system, we have implemented a PIM module that can perform vector operations with different operand sizes using the proposed set of tools.

Full Paper:

processing-in-memory | near-data processing | simulator | hybrid memory cube | systemc

Research Talks

ICDE 2022

Polynesia: Hybrid Transactional/Analytical DBs

26 mins | Video: | Slides:

ISVLSI 2022

ISVLSI 2022 Special Session on Processing-in-Memory

33 mins | Video: | Slides:

P&S PIM, ETH Zurich

SIMDRAM: An End-to-End Framework for Bit-Serial SIMD Computing in DRAM

1 hour 11 mins | Video: | Slides:

Service

2023
Reviewer for IEEE Transactions on Computers
Subreviewer for ASPLOS, MICRO, EuroSys

2022
Reviewer for IEEE Transactions on Computers
Subreviewer for ASPLOS, DSN, ISCA, CAL, TCAD, and USENIX ATC

2021
Subreviewer for HPCA, MICRO, TCAD, and USENIX ATC

2020
Subreviewer for DSN, ISCA, MICRO, CCS, ISCAS, ISPASS, NVMW, and TCSI

2019
Subreviewer for DSN, ISCA, MICRO, MSST, TCAD, and TED

2018
Subreviewer for ASPLOS, HPCA, PACT, Nature Electronics, TC, and TVLSI

2017
Subreviewer for DSN, MICRO, ISCA, and PLDI

Employment

Nov 2018 - Present
Research and Teaching Assistant - ETH Zürich
Apr 2019 - Sep 2019
Hardware Engineering Intern - Google
Nov 2017 - Oct 2018
Research Intern - ETH Zurich
Mar 2016 - Aug 2017
Research and Teaching Assistant - Federal University of Rio Grande do Sul
Jul 2015 - Dec 2015
Undergrad Research Assistant - Federal University of Vicosa
Mar 2014 - May 2014
Undergrad Research Assistant - Federal University of Vicosa

Other Publications

J. Gómez-Luna, Y. Guo, S. Brocard, J. Legriel, R. Cimadomo, G. F. Oliveira , G. Singh, O. Mutlu **"Evaluating Machine Learning Workloads on Memory-Centric Computing Systems"** Proceedings of the 2023 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Raleigh, North Carolina, USA, April 2023. Full Paper: Full Talk (15 mins):

M. Item, J. Gómez Luna, Y. Guo, G. F. Oliveira , M. Sadrosadati, O. Mutlu, **"TransPimLib: Efficient Transcendental Functions for Processing-in-Memory Systems"** Proceedings of the 2023 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Raleigh, North Carolina, USA, April 2023. Full Paper: Full Talk (17 mins):

J. D. Ferreira, G. Falcao, J. Gómez-Luna, M. Alser, L. Orosa, M. Sadrosadati, J. S. Kim, G. F. Oliveira , T. Shahroodi, A. Nori, O. Mutlu **"pLUTo: Enabling Massively Parallel Computation in DRAM via Lookup Tables"** Proceedings of the 55th International Symposium on Microarchitecture (MICRO), Chicago, IL, USA, October 2022. Full Paper: Lecture Video (26 mins):

J. Park, R. Azizi, >G. F. Oliveira , M. Sadrosadati, R. Nadig, D. Novo, J. Gómez-Luna, M. Kim, O. Mutlu, **"Flash-Cosmos: In-Flash Bulk Bitwise Operations Using Inherent Computation Capability of NAND Flash Memory"** Proceedings of the 55th International Symposium on Microarchitecture (MICRO), Chicago, IL, USA, October 2022. Full Paper: Lecture Video (44 mins):

A. G. Yağlıkçı, H. Luo, G. F. Oliveira , A. Olgun, M. Patel, J. Park, H. Hassan, J. S. Kim, L. Orosa, O. Mutlu **"Understanding RowHammer Under Reduced Wordline Voltage: An Experimental Study Using Real DRAM Devices"** Proceedings of the 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Baltimore, MD, USA, June 2022. Full Paper: Full Talk (34 mins):

A. Boroumand, S. Ghose, G. F. Oliveira , O. Mutlu, **"Polynesia: Enabling High-Performance and Energy-Efficient Hybrid Transactional/Analytical Databases with Hardware/Software Co-Design,"** in Proceedings of the 38th International Conference on Data Engineering (ICDE), Virtual, May 2022. Full Paper: Full Talk (25 mins):

M. Patel, G. F. Oliveira , O. Mutlu, **"HARP: Practically and Effectively Identifying Uncorrectable Errors in Memory Chips That Use On-Die Error-Correcting Codes,"** in Proceedings of the 54th International Symposium on Microarchitecture (MICRO), Virtual, October 2021. Full Paper: Full Talk (20 mins):

A. Boroumand, S. Ghose, B. Akin, R. Narayanaswami, G. F. Oliveira , X. Ma, E. Shiu, O. Mutlu, **"Google Neural Network Models for Edge Devices: Analyzing and Mitigating Machine Learning Inference Bottlenecks,"** in Proceedings of the 30th

International Conference on Parallel Architectures and Compilation Techniques (PACT), Virtual, September 2021. Full Paper: Full Talk (14 mins):

N. Hajinazar, P. Patel, M. Patel, K. Kanellopoulos, S. Ghose, R. Ausavarungnirun, G. F. Oliveira , J. Appavoo, V. Seshadri, O. Mutlu, **"The Virtual Block Interface: A Flexible Alternative to the Conventional Virtual Memory Framework,"** in Proceedings of the 47th International Symposium on Computer Architecture (ISCA), Valencia, Spain, June 2020. Full Paper: Full Talk (26 mins):

G. Singh, J. Gómez-Luna, G. Mariani, G. F. Oliveira , S. Corda, S. Stuijk, O. Mutlu, H. Corporaal, **"NAPEL: Near-Memory Computing Application Performance Prediction via Ensemble Learning,"** in Proceedings of the 56th Design Automation Conference (DAC), Las Vegas, NV, USA, June 2019. Full Paper:

P. C. Santos, G. F. Oliveira , D. G. Tomé, M. A. Z. Alves, E. C. Almeida and L. Carro, **"Operand Size Reconfiguration for Big Data Processing in Memory,"** in Proceedings of the 2017 Design, Automation & Test in Europe Conference & Exhibition (DATE), Switzerland, March 2017. Full Paper:

Last update: July 25, 2023