Vanderson Martins do Rosario

Senior GPU Compiler Engineer at NVIDIA

San Jose, California | vandersonmr2@gmail.com | GitHub | LinkedIn | Google Scholar

I build compilers for GPUs and specialized accelerators. My research includes operator fusion, kernel scheduling, performance prediction, and dynamic compilation.

Professional experience

NVIDIA 2026-present

Senior GPU Compiler Engineer

Work on GPU backend compilation and code generation.

Cadence Design Systems 2020-2026

Software Architect, previously Senior Principal and Principal Software Engineer

Led a nine-engineer team developing the MLIR-based Xtensa Neural Network Compiler. Designed compiler architecture and NPU code generation, developed LLVM optimization passes, and proposed integration paths for PyTorch via torch-mlir and for StableHLO.

Microsoft Research 2019

Research Intern, Cambridge, United Kingdom

Implemented rematerialization in TensorFlow/XLA to reduce accelerator memory pressure. The evaluation measured 69% lower memory consumption against the baseline with less than 11% average runtime overhead.

MATERA Systems 2015-2016

Research and Development Intern, Maringá, Brazil

Worked on compiler tools for migrating legacy software.

Education

University of Campinas (Unicamp) 2016-2021

Ph.D. in Computer Science

Dissertation: *Accelerating Capsule Networks with Lanes*. Best doctoral thesis, WSCAD-CTD 2022.

Instituto Superior Técnico / INESC-ID 2018-2019

Visiting Ph.D. student, Lisbon, Portugal

Studied neural-network parallelism and mapping neural-network lanes to hardware.

State University of Maringá (UEM) 2011-2015

B.S. in Computer Science

Research & open source

Publication records | Google Scholar | DBLP

Selected publications

Fusion of Operators of Computational Graphs via Greedy Clustering: The XNNC Experience 2025

Michael Canesche, Vanderson Martins do Rosario, Edson Borin, and Fernando Magno Quintão Pereira.

ACM SIGPLAN International Conference on Compiler Construction

The Droplet Search Algorithm for Kernel Scheduling 2024

Michael Canesche, Vanderson Martins do Rosario, Edson Borin, and Fernando Magno Quintão Pereira.

ACM Transactions on Architecture and Code Optimization

Fast Selection of Compiler Optimizations Using Performance Prediction with Graph Neural Networks 2023

Vanderson Martins do Rosario, Anderson Faustino da Silva, André Felipe Zanella, Otávio O. Napoli, and Edson Borin.

Concurrency and Computation: Practice and Experience

Smart Selection of Optimizations in Dynamic Compilers 2021

Concurrency and Computation: Practice and Experience

Selected software and research programs

TensorFlow/XLA rematerialization · Compiler pass for reducing accelerator memory use.

Code Performance Predictor · Graph neural network model for performance prediction.

Multi-Lane Capsule Networks · Research code for separable capsule networks.

Google Summer of Code, QEMU (2019) · Built tools to inspect TCG code generation and identify emulation anomalies.

Samsung research fellowship (2017-2018) · Led algorithm development for a static analysis tool that used JIT profiles to detect Android energy bugs.

Technical areas

GPU and NPU code generation, MLIR and LLVM, TensorFlow/XLA, QEMU TCG, C++, and Python.