Elliott Binder

ebinder@cmu.edu | (814) 203 0173 | Pittsburgh, PA

PhD Candidate and researcher specializing in the modeling and design of high-performance computation for CPUs, GPUs, FPGAs, and NPUs, from single-core kernels to distributed processing. Proven expertise in identifying performance bottlenecks and developing novel algorithms and analytical models to accelerate complex computational tasks in machine learning, linear algebra, and signal processing.

Work Experience

Carnegie Mellon University — Pittsburgh, PA
Research Assistant — 09/2017 - Present

- Analyzes computation and data dependencies, guiding the design of efficient hardware and software implementations.
- Develops performance models for architectures (CPU, GPU, NPU, FPGA) based on hardware capabilities.
- Evaluates algorithm and implementation designs against state-of-the-art solutions and theoretical peak.

AMD Research — Remote
Research Intern — 05/2021 - 08/2021

- Investigated the performance benefits and architectural implications of potential future vector extensions for CPUs.

AMD — Remote
MIOpen Intern — 05/2020 - 08/2020

- Developed a detailed performance model for CNNs on GPUs and optimized index calcluations for implicit GEMM.

Los Alamos National Laboratory — Los Alamos, NM
05/2018 - 08/2018

- Completed intensive coursework in network archaeology, host forensics, and malware analysis.
- Led an incident response project analyzing network and host data, presenting findings to senior management.

IBM — Poughkeepsie, NY
Millicode Firmware Engineer — 06/2016 - 08/2017

- Developed and maintained embedded software for IBM z/Systems mainframes.
- Analyzed and debugged complex hardware traces from both field and test environments.
- Designed and implemented internal tools to expedite development and debugging processes.

Education

Carnegie Mellon University — Pittsburgh, PA
M.S./Ph.D. , Electrical and Computer Engineering | GPA: 3.92 — 12/2025

Johns Hopkins University — Baltimore, MD
B.S. , Computer Engineering | GPA: 3.55 — 05/2016

Projects

Fast Attention Through Optimizing Memory 06/2025
Designed specialized algorithms and kernels for the memory-bound matrix multiplications found in multi-head attention layers, differing from canonical HPC implementations to optimize for large, non-standard input sizes (presented at IPDPS 2025, published in proceedings).

Architecture-Aware Models of AI Engines for High-Performance MMM 09/2025
Extend analytical models to the AIE-ML architecture to guide the design of high performance matrix multiplication kernels at the register, compute tile, and whole array levels, achieving greater than 90% of compute peak (to be presented at ICPP 2025 and published in proceedings).

Enabling Memory Optimizations for Matrix Acceleration in MLIR 12/2025
Developing methods for improving efficiency of codes leveraging matrix acceleration instructions on CPUs and GPUs using MLIR.

Scalable On-Array Processing of Radar Systems Using Machine Learning 07/2025
Distributed system design and evaluation of sub-quadratic scaling algorithms for beamforming and ML-based detection pipelines.

Portable Implementations of Prime-Sized Discrete Trigonometric Transforms 05/2024
Designed a novel O(N^2) algorithm that outperforms state-of-the-art O(N log N) libraries by exploiting matrix symmetry, demonstrating portability across CPU/GPU vendors and various transform types.

Distributed Multi-Dimensional Fast Fourier Transforms 09/2023
Developed a communication-aware decomposition for distributed MDFFTs, mapping computation at the node, GPU, block, and thread levels to minimize data movement, reduce latency, and maximize throughput.

Analytical Model for Portable Performance of MMM-like Operations on GPUs 11/2022
Derived software parameters directly from hardware features to produce high-performance, portable implementations for matrix-multiply algorithms across diverse GPU architectures.

A Portable GPU Framework for SNP Comparisons 05/2019
Developed a framework for analyzing single nucleotide polymorphisms (SNPs) to detect the presence/absence of minor alleles, portable to a variety of GPU platforms.

Skills

Languages & Tools:
C, C++, Python, Assembly, CUDA, HIP, SYCL, OpenCL, OpenMP, MPI, HLS, gem5, LLVM, MLIR

Expertise:
Deep Learning, Signal Processing, High Performance Computing, Algorithm Design & Optimization, Performance Modeling, Computer Architecture, Embedded Systems, Dense Linear Algebra