Alessandro Carraro

Edinburgh, UK alecarraro33@gmail.com alecarraro alecarraro.github.io

EDUCATION

University of Edinburgh September 2026 – Present
PhD in Computational Mathematics
Research: AI Model Compression via Structured Linear Algebra

KTH Royal Institute of Technology & Delft University of Technology September 2024 – August 2026
M.Sc. in Applied Mathematics (Joint Degree)
Cum Laude, GPA: 8.99/10 – Top 3% among Applied Mathematics graduates at TU Delft

Delft University of Technology – B.Sc. in Aerospace Engineering September 2021 – June 2024

EXPERIENCE

Google Summer of Code Contributor, Julia Lab, MIT – `NextLA.jl` May 2026 – August 2026
Project: Tile-Low-Rank GEMM on GPUs

- Brought Tile-Low-Rank linear algebra to GPUs: compression, GEMM with dense and compressed accumulation, and TLR variants such as dense-diagonal matrices
- Co-designed the data structure with the GEMM lowering, fusing many small tile updates into a few large **zero-copy GEMMs** with exact per-tile ranks and a bounded workspace
- On H100: **2.6× faster than KBLAS**, reaching **92% of dense cuBLAS GEMM throughput**

Google Summer of Code Contributor, JuliaReach May 2025 – August 2025
Reachability Analysis for Control of Uncertain Dynamical Systems

- Implemented recent reachability algorithms in the open-source `JuliaReach` ecosystem; co-authored the ARCH-COMP 2026 report

Graduate Teaching Assistant, TU Delft September 2024 – March 2025
High Performance Computing labs (M.Sc. and B.Sc.): MPI, OpenMP, and CUDA

RESEARCH

***New variant* of Block Jacobi SVD on Hopper GPUs** September 2026 – Present
Paper in preparation; preprint expected October 2026

- Designed a Jacobi SVD variant needing about **2× fewer FLOPs** to converge, with high relative accuracy
- Custom kernels: batched tall-and-skinny SYRK on CuTe FP64 MMA atoms with TMA loads, and a shared-memory-resident batched eigensolver reformulated around GEMMs with a custom swizzled layout
- On H200: up to **5.0× faster than cuSOLVER** `gesvd`, 9.9× than `gesvdj`; beats `gesvdp` to $n = 2048$

A GPU-Accelerated Mixed-Precision Block-Jacobi SVD (M.Sc. thesis) January 2026 – June 2026

- Designed a mixed-precision one-sided block Jacobi SVD in which cheap FP16, TF32 or FP32 sweeps precondition the matrix before FP64 refinement, with an accuracy-aware criterion.
- Up to **4.7× faster** than pure FP64 on L40S at FP64 accuracy; no gain on H100, where FP64 uses DMMA

SKILLS

Programming Languages: C++, Julia, Python
GPU Libraries: CUDA C++, CuTe, CUTLASS (basic), cuBLASDx
Tools: Nsight Compute, Nsight Systems, Git, LaTeX
Languages: Italian (native), English (fluent)