

Truls E. Stokke

SOFTWARE ENGINEER

CONTACT

@ contact@truls.es web truls.es gh [crypt3lx2k](https://github.com/crypt3lx2k) in [LinkedIn](https://www.linkedin.com/in/truls-es)

SUMMARY

Software Engineer with a background in HPC and AI Infrastructure. My HPC work was primarily distributed computing with MPI and low-level compute optimization using SIMD instruction sets. In AI I have worked mainly with non-standard accelerator ASICs, doing end-to-end work all the way from low-level on-core assembly programming to serving-stack optimization.

EXPERIENCE

Meta Platforms, Inc. | *Software Engineer*

MAR 2023 - OCT 2026

- Worked in the AI Applications Engineering team on Meta's in-house AI accelerator (MTIA).
- Improved on-chip memory bandwidth by up to 10% for element-wise operations by exploiting the physical layout of the Network-on-Chip.
- Enabled a new class of models for a Product Group, yielding a 40% improvement in Performance/\$ over existing GPU systems.
- Wrote a class of non-ML compute kernels for intersection of unsorted integer streams using RISC-V SIMD instructions, for a 100% improvement in Performance/\$ over existing GPU systems.
- Enabled a new feature to optimize sparse data movement between host and device, from firmware level up through the PyTorch stack.

Graphcore Limited | *AI Engineer*

JAN 2020 - JAN 2023

- Worked primarily in the Applications team as a customer-facing engineer.
- Ported a Partial Differential Equation solver to the Graphcore IPU for a customer, achieving a throughput of 4+ TB/s.
- Worked on a fine-grained dynamic sparsity project for a customer, enabling the RigL method for LLM training.
- Optimized the IPU inference serving stack for a key customer, cutting P99 latency 10x through kernel and network configuration changes.

PUBLICATIONS

MTIA 300: Meta's First Training Chip Featuring Built-in NICs and Collective Offloading Engines

C. Tang et al. (MTIA Team) | ISCA 2026, ACM/IEEE 53rd Annual International Symposium on Computer Architecture

Meta's Second Generation AI Chip: Model-Chip Co-Design and Productionization Experiences

J. Coburn, C. Tang, et al. | ISCA 2025, 52nd Annual International Symposium on Computer Architecture

EDUCATION

Bachelor of Science, Applied Mathematics

University of Oslo

HPC focus: numerical methods, performance, and distributed computing.

SKILLS

LANG

Python, C/C++, x86 AVX Assembly, RISC-V RVV Assembly

HPC

CUDA, MPI, OpenMP

AI

PyTorch, TensorFlow

OTHER

Docker, Kubernetes