

Kumudha Narasimhan

✉ kumudhakn@gmail.com

📍 United Kingdom

🔗 kumudhan.github.io/

🌐 in/kumudha-narasimhan

SUMMARY

Technical and people leader with 10 years of domain experience in AI and HPC performance optimization across diverse architectures with multiple research publications. Elevates team performance by engaging with technical experts and guiding contributions toward successful outcomes, while aligning with overall business strategy. Experienced in mentoring, coaching and collaborating with engineers and leaders at all levels of the organization. Creates an inclusive environment where everyone can thrive.

SKILLS

Performance optimization

oneDNN, oneMKL, ONNX Runtime and using cuBLAS, cuDNN, CUTLASS, llama.cpp

Research projects Managed

EU projects - AERO (open cloud acceleration), SYCLOPS (AI acceleration)

Programming Paradigms

SYCL, CUDA, C++, Python

Leadership & Process

Agile methodologies, roadmap design, OKR alignment, process audit, mentoring

EXPERIENCE

Jan 2023 – Present
Edinburgh, UK

Software Engineering Manager

Intel Subsidiary - Codeplay Software

- Managed 4 specialized Agile teams (18+ engineers) optimizing LLMs - llama.cpp for Intel GPUs and developing SYCL backend for CUTLASS and Triton.
- Led evaluation of LLM accuracy and correctness for llama.cpp alongside performance optimization, designing benchmarking frameworks to validate reliability and reproducibility.
- Defined evaluation criteria to detect errors and hallucinations during optimization pipelines, ensuring model consistency across platforms.
- Oversaw EU projects (AERO, SYCLOPS), aligning project goals with organizational priorities.
- Provided people-focused leadership with regular 1:1s, performance reviews, onboarding, and HR issues across internal transitions.
- Collaborated across HR, infrastructure, and leadership teams to enhance organizational workflows and improve team health.

Nov 2020 – Dec 2022
Edinburgh, UK

Senior Staff Engineer

Codeplay Software

- Led SYCL integration efforts for projects like ONNX Runtime and implemented features in libraries such as portBLAS, portDNN and portFFT to optimize performance on heterogeneous hardware.
- Built evaluation pipelines and CI workflows to improve reproducibility and systematically track correctness of results.
- Published multiple peer-reviewed papers (PMAM 2022, TACO 2023, P3HPC 2022) by driving technical content, managing review/rebuttal processes, and presenting results at major conferences and client demos.
- Contributed to error analysis and validation in performance portability experiments, ensuring reproducibility and correctness across architectures.

Sep 2019 – Nov 2020
Edinburgh, UK

Staff Software Engineer

Codeplay Software

- Contributed to adding cuBLAS and cuDNN backends into oneMKL and oneDNN to enable GPU acceleration portability for these libraries. Collaborated with compiler teams to develop SYCL features to enable SYCL as a backend for ONNX Runtime.
- Optimized neural network inference on custom hardware for real-time performance as part of a customer-facing project.
- *Published research* on deep learning performance optimization in top-tier venues (ICS 2021, P3HPC 2020) - tile size selection model and cross- platform DNN portability.
- *Presented SYCL tutorials* and major events, including IWOCL 2021 and Intel Dev Summit 2021.

Jul 2018 – Jul 2019
Bengaluru, India

Senior Software Engineer

Samsung Research India

- Analyzed *custom profilers and debuggers* for Samsung Tizen native and web applications, enhancing their functionality based on IoT team requests.
- Investigated and resolved issues reported in the Tizen Studio IDE to improve developer experience.
- Coached colleagues for the professional-level internal competitive coding exam, focusing on data structures and algorithms.

Jun 2012 – Jul 2015
Bengaluru, India

Technology Analyst at Investment Banking Division

Goldman Sachs

- Building and standardizing frameworks in C# .NET and Java for all applications developed within the division with *focus on scalability and performance*.
- Inherited several C# projects and led their sustenance, ongoing development, and enhancements, including re-architecting the resource discovery system, entitlement system, and provisioning system.
- Communicated effectively with internal and business stakeholders, inducted and mentored new joiners, and resolved multiple critical issues within deadlines while managing expectations.

EDUCATION

2018
Bangalore, India

M.Sc. (by research), High Performance Computing & Compilers

Indian Institute of Science (Advisor: Prof. Uday Reddy)

Thesis: Optimizing dense matrix computations with PolyMage

2012
Bangalore, India

Bachelor of Engineering, Computer Science and Engineering

M S Ramaiah Institute of Technology

Selected Publications

- Towards performance portability of AI graphs using SYCL. [P3HPC@SC 2022]
- A practical tile size selection model for affine loop nests. [ICS 2021]
- Towards Cross-Platform Performance Portability of DNN Models using SYCL. [P3HPC@SC 2020]
- Optimizing geometric multigrid method computation using a DSL approach [SC 2017]

Full list of Publications - <https://dblp.org/pid/208/1873.html> ↗

List of presentation / talks - <https://kumudhan.github.io/#publications> ↗