Kumudha Narasimhan

kumudhakn@gmail.com
United Kingdom
in/kumudha-narasimhan
https://kumudhan.github.io/

SUMMARY

Technical program and Engineering Manager with 10+ years of experience leading complex software and research programs in AI, recently working on LLMs. Experienced in working with researchers and engineers to build prototypes, define and execute software technology roadmaps, manage project schedules, dependencies, risks and stakeholder alignment to deliver high-impact solutions in partnership with teams. Skilled at leading diverse engineering teams, guiding technical direction, and aligning research outcomes with product and strategic goals.

SKILLS

AI
LLM evaluation and benchmarking, Prompt design & correctness testing, Hallucination analysis, Model reliability & cross-platform consistency, Rapid prototyping (Python, notebooks)

Leadership and Product
Technology roadmapping, OKR creation and tracking, ISO process audit, Mentoring, Agile methodologies, Risk and dependency management, Budget and resource planning, Technical storytelling

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

Research projects Managed
EU projects - AERO (open cloud acceleration), SYCLOPS (Al acceleration)

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.
- Transformed technical findings into clear narratives and demos for internal leadership and product stakeholders.
- Oversaw EU projects (AERO, SYCLOPS), aligning project goals with organizational priorities. Created prototypes and demos in collaboration with other members for project milestones.
- Provided people-focused leadership with and collaborated across HR, infrastructure, and leadership teams to enhance organizational workflows and improve team health

Nov 2020 – Dec 2022
Edinburgh, UK

Senior Staff Engineer (Tech Lead)
Codeplay Software

- Managed SYCL integration in ONNX Runtime project to enable portability of AI applications. Developed simple and visual demos for conference presentations and customer engagements.
- Contributed to error analysis and validation in performance portability experiments, ensuring reproducibility and correctness across architectures.
- Co-authored multiple peer-reviewed papers by driving technical content, managing review/rebuttal processes, and presenting at major conferences and client demos.

- Improved engineering workflows and technical processes by establishing CI pipelines, improved benchmark reproducibility, defining acceptance criteria for MRs and supporting ISO audit readiness through systematic best practices.

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 portability for these libraries.
- 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 contributing to early stage prototypes to enable cross-platform DNN execution.
- Presented SYCL tutorials and major events, including IWOCL 2021 and Intel DevSummit 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 Research Partnerships

Co-authored multiple publications with research collaborators, contributing to experiment design, prototype evaluation, and cross-functional knowledge sharing.

- 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]
- The oneAPI Software Abstraction for Accelerated Computing [Panel at Supercomputing (SC), Nov 2022]

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

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