Kumudha Narasimhan

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

SUMMARY

Technical and people leader with 10 years of experience in building and scaling high performance software systems across diverse architectures. Skilled in defining and executing software technology roadmaps, driving productization of research into production-ready solutions, and ensuring quality and integration maturity. Proven ability to engage with customers, open-source communities, and industry partners to gather requirements, influence adoption, and deliver high-impact solutions. Experienced in mentoring and managing engineering teams while aligning technical execution with business strategy.

SKILLS

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

Programming Paradigms
SYCL, CUDA, C++, Python

Leadership & Process
Agile methodologies, Technology roadmapping, OKR alignment, ISO process audit, mentoring, requirements prioritization, competitive analysis

EXPERIENCE

Jan 2023 – Present
Edinburgh, UK

Software Engineering Manager
Intel Subsidiary - Codeplay Software

- Provide people-focused leadership to 18+ engineers across 4 Agile teams, including senior and principal engineers and tech leads. The teams focused on
  - Compiler implementation and optimizations to enhance FP16 matrix multiplication performance for the Intel XPU backend of the Triton compiler.
  - SYCL backend support for CUTLASS library
  - Optimized llama.cpp implementation for Intel devices using SYCL.
  - Collboration on 2 EU projects(AERO, SYCLOPS) to enable SYCL applications on RISC-V vector processors and ARM architectures.
- Defined and executed roadmaps for all teams, ensuring alignment with Intel's product strategy.
- Partnered with cross-functional stakeholders (HR, Agile coaches, PMs, infrastructure teams, leadership) to improve organizational processes, track OKRs, and guide initiatives.

Nov 2020 – Dec 2022
Edinburgh, UK

Senior Staff Engineer (Tech Lead)
Codeplay Software

- Developed a user-driven kernel fusion approach by extending the SYCL API to automate short-running kernel fusion, reducing overhead in graph-based neural networks while preserving modularity and portability.
- Developed a CPU-directed compilation flow for SYCL, bypassing the OpenCL backend to eliminate redundant overhead and integrate Whole Function Vectorization, achieving improved performance of SYCL applications on CPUs without code modifications.
- Enhanced engineering maturity by defining CI pipelines, reproducibility standards, and merge review protocols, contributing to ISO audit readiness.
- Represented company in external communities by co-authoring peer-reviewed publications, presenting at conferences (PMAM, TACO, P3HPC), and delivering customer demos.

Sep 2019 – Nov 2020
Edinburgh, UK

Staff Software Engineer
Codeplay Software

- Developed GPU backends (cuBLAS, cuDNN) for oneMKL and oneDNN, increasing portability and adoption across hardware platforms.
- Enabled SYCL as a backend for ONNX Runtime, strengthening ecosystem engagement.
- Optimized neural network inference using glow compiler on custom hardware for real-time performance as part of a customer-facing project.
- Presented technical tutorials at major events (IWOCL, Intel Dev Summit), evangelizing SYCL and performance portability to the wider community.

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 and Talks

- Programming Model Extensions for General-Purpose Processing-In-Memory. [ISC 2024]
- Accelerating Neural Networks Using Open Standard Software on RISC-V. [International workshop@ISC 2023]
- Improving performance of SYCL applications on CPU architectures using LLVM-directed compilation flow [PMAM@PPoPP 2022]
- User-Driven Online Kernel Fusion for SYCL [TACO 2022]
- A practical tile size selection model for affine loop nests. [ICS 2021]
- Towards performance portability of AI graphs using SYCL. [P3HPC@SC 2022]

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

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