Sourav Chakraborty

Austin, TX, USA (614) 500-3409 mail@souravc.com

SUMMARY

Over a decade of experience in high-performance computing (HPC), AI infrastructure, and LLM finetuning and inference. Expertise in hardware software co-design, developing LLM-powered full-stack web applications with responsive frontend and scalable backend. Proven track record in leading cross-functional, globally distributed teams to deliver cutting-edge software solutions. Passionate about leveraging AI to build highly innovative and user-centric solutions.

EXPERIENCE (HPC, AI, HW-SW Codesign)

SAMSUNG - ***Senior Staff Software Engineer*** **JAN 2023 - PRESENT**

- Architect communication runtime for a novel supercomputer system designed for HPC applications and AI workloads including LLM training and inference
- Lead a globally distributed team of developers and researchers to envision, plan, and execute development roadmap
- Lead research on modeling, performance projection, and competitive analysis of AI applications on new hardware
- Collaborate with hardware team to provide functional and performance requirements

NVIDIA - ***HPC Middleware Developer*** **JUN 2021 - JAN 2023**

- Lead development of offloaded collective framework for NVIDIA BlueField DPUs to accelerate HPC applications and AI frameworks like PyTorch
- Contribute to strategic planning and execution of development roadmap for NVIDIA SmartNICs software platform
- Design collective algorithms to maximize performance and overlap of communication and computation with limited system resources
- Collaborate with inside sales teams and external customers to identify, prioritize, and quickly solve customer requirements for developing AI software and deploying large AI clusters

AMD - ***Senior Software Engineer*** **OCT 2019 - JUN 2023**

- Lead UCX and MPI development efforts for AMD GPUs using ROCm platform
- Improve performance of point-to-point (2x) and collective operations (8x) on AMD GPUs on RDMA capable networks like InfiniBand and RoCE
- Collaborate with DevOps team to set up functionality and performance QA process for UCX, MPI, and RCCL releases
- Identify and contribute support for AMD GPUs to strategically relevant open-source software ecosystems (RDMA, MPI Benchmarks, etc.)
- Helped secure $5M+ deal by analyzing and quickly solving customer need

Ohio State University - ***Graduate Research Assistant*** **AUG 2013 - AUG 2019**

- Develop scalable and adaptive communication runtimes for large-scale HPC and AI systems
- Design intra-node and inter-node communication protocols and collective algorithms for MPI and OpenSHMEM on CPU and GPU clusters with InfiniBand and OmniPath interconnects
- Develop optimized MPI libraries for cloud environments including AWS and Azure
- Develop scalable fault-tolerant MPI library with SLURM

EXPERIENCE (Full-Stack Web Development, LLM Integration)

Hiremator.com - ***AI Powered Recruitment Platform*** **2023 - 2024**

- Developed responsive frontend using React.js and Next.js, creating a dynamic user interface
- Designed and developed a robust and performant backend with Next.js, and PostgreSQL
- Integrated with OpenAI GPT APIs to streamline recruitment functionalities, including resume parsing and screening, intelligent candidate matching and candidate assessments

Didimoni.com - ***Whatsapp based AI Tutor*** **2023 - 2024**

- Utilized OpenAI GPT APIs to provide real-time, AI-powered tutoring for Indian students, facilitating interactive and personalized learning experiences
- Designed and developed a webhook based backend with Python and FastAPI
- Used retrieval augmented generation (RAG) and vector databases to index books, syllabus, and other academic content and provide contextual answers to students

Yahoo! - ***Software Development Engineer*** **JUL 2011 - AUG 2013**

- Develop frontend components and backend APIs for storing and serving ~1 Billion user profiles to high-traffic pages including Yahoo! Frontpage
- Develop services to import ~500M Facebook profiles to Yahoo! NoSQL database

EDUCATION

Ohio State University - ***MS, PhD, Computer Science and Engineering*** **2013-2019**

Thesis: High Performance and Scalable Cooperative Communication Middleware for Next Generation Architectures

Jadavpur University - ***BE, Information Technology*** **2007-2011**

TECHNOLOGIES

AI, LLM, RDMA, MPI, UCX, GPUDirect, InfiniBand, SHARP, NCCL, SLURM, Python, Javascript, Typescript, React.js, Next.js, Prisma ORM, SQL, NoSQL, RAG, Vector DB, LLM Fine-tuning

PUBLICATIONS AND PATENTS

25+ peer-reviewed papers in journals and conferences. https://souravc.com/pubs/
US20140012906A1: Peer-to-peer architecture for web traffic management (Yahoo!)
US20240095062A1: Offloaded task computation on network-attached co-processors (Nvidia)

AWARDS

Best Graduate Student Research Award, OSU-CSE 2018
ACM Student Research Award at SuperComputing 2016
Several **best paper and best poster awards** and nominations at conferences including SC, IPDPS, ISC, and CLUSTER
Represented employers in MPI Forum and other standardization bodies