Sourav Chakraborty

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

SUMMARY

Over a decade of leadership experience in high-performance computing (HPC), AI infrastructure, and large language models (LLM). Proven track record in driving strategic AI vision and shaping product roadmaps that balance technical innovation with business goals. Expertise in hardware-software co-design, delivering cloud-native generative AI deployments, and spearheading collaborations with major global partners like Meta and Microsoft. A visionary founder who has led cross-functional teams and designed AI-driven, user-centric solutions for clients, with a focus on strategic innovation, rapid prototyping and delivery of business value.

EXPERIENCE

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

- Shaped the strategic vision for a next-generation supercomputing project for AI and HPC, directly influencing both technical and business outcomes.
- Led a globally distributed team of engineers and researchers, driving research and development roadmaps that aligned with broad industry trends, market research, and customer needs.
- Designed a novel high-performance communication runtime, leading to more than 2x projected performance improvement in HPC and AI workload efficiency.
- Delivered multi-million-dollar value to the business by leading initiatives that anticipated future AI infrastructure needs, ensuring Samsung's leadership in the AI and HPC markets.

NVIDIA - ***Staff Software Engineer*** **JUN 2021 - JAN 2023**

- Led the development of offloaded collective frameworks for NVIDIA BlueField DPUs, contributing to 50% faster AI and HPC data communication, accelerating AI training times for enterprise-level clients.
- Defined and executed long-term strategy for NVIDIA SmartNICs, ensuring that middleware advancements translated into a 30% performance boost for AI-heavy HPC applications.
- Collaborated with external clients to rapidly address AI workload requirements, enabling new AI-driven use cases and enhancing NVIDIA's market position in AI networking.

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

- Delivered 2x improvements in point-to-point operations and 8x improvements in collective operations on AMD GPUs by driving innovation in communication libraries for RDMA-capable networks.
- Optimized AMD's GPU performance on InfiniBand, directly contributing to a $5M+ deal by quickly solving critical customer needs related to AI workload performance.
- Championed collaboration between AMD's hardware and software teams to deliver scalable, fault-tolerant AI solutions, securing AMD's position as a key player in AI acceleration technologies.
- Identified and contributed support for AMD GPUs to strategically relevant open-source software ecosystems (RDMA, MPI Benchmarks, etc.)

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

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

SIDE PROJECTS

Hiremator - ***AI Powered Recruitment Platform***

- Developed an innovative AI-powered recruitment platform using OpenAI GPT-powered APIs for 90% faster resume parsing and candidate matching.
- Led end-to-end development using React.js, Next.js, and PostgreSQL to build a scalable platform, resulting in improved recruitment accuracy and efficiency for clients.

DidiMoni - ***WhatsApp-based AI Tutor***

- Designed and developed an AI-powered tutoring platform, delivering personalized educational assistance to Indian students through WhatsApp using GPT APIs and RAG.

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**

Project: Extractive Text Summarization using K-means Clustering and TF-IDF

TECHNOLOGIES

AI, LLM, RDMA, MPI, UCX, GPUDirect, InfiniBand, SHARP, NCCL, PyTorch, Python, JavaScript, TypeScript, React.js, Next.js, Prisma ORM, AWS, Azure, RAG, Vector DB, LLM Fine-tuning, NLP

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

Outstanding Graduate Student Research Award - Ohio State University
First Place, ACM Student Research Award - International Conference of SuperComputing
Several **best paper and best poster awards** at conferences including SC, IPDPS, ISC, and CLUSTER