Richi Dubey

Atlanta, GA 30318 | (470)3387465 | richidubey@gatech.edu | linkedin.com/in/richidubey/ | github.com/richidubey | https://richidubey.github.io/blog/ (Blog)

EDUCATION

Georgia Institute of Technology – Atlanta, GA *August 2024 - June 2026 (Expected)*
Masters, Computer Science (MS CS : Machine Learning Specialization), GPA 3.9/4
Relevant coursework: Deep Learning, Computer Vision, Natural Language, High Performance Computing

Birla Institute of Technology & Science, Pilani – Goa, India *Aug 2017 - June 2021*
Bachelors, Computer Science (BS CS), GPA: 3.9/4
Relevant coursework: Real-Time Systems, Data Structures & Algorithms, Operating Systems (A grade)

WORK EXPERIENCE & INTERNSHIPS

NVIDIA – Santa Clara, CA
Software Engineer Intern, CUDA Driver Team *May 2025 – August 2025*
- Implemented support for parameterized CUDA Graphs, enhancing composability and reducing node update overhead, resulting in **up to 24%** improvement in ML workload performance.

CERN (European Organization for Nuclear Research) – Geneva, Switzerland
Software Engineer *October 2022 – July 2024*
- Built high-performance, multithreaded C++ drivers for a **distributed** SCADA **system monitoring 1000+** sensors across CERN accelerator facilities
- Developed fault-tolerant networking modules, maintaining downtime of less than 0.01%

Oracle – Bangalore, India
Software Engineer *July 2021 – September 2022*
- Implemented **4+ production features** in a cloud microservices application built with Java Spring Boot.
- Automated global deployments with Terraform/Kubernetes across **30+ OCI data centers**, cutting deployment effort for the team by **30%**.

OPEN SOURCE CONTRIBUTIONS

RTEMS Real Time Operating System (RTOS) – Google Summer of Code *Summer 2020*
- Contributed 1000+ lines of code to RTEMS for implementing a custom scheduler, improving system schedulability by **20%**. Work published at **EMSOFT 2021** (ACM) and wrote a blog on the implementation.

PUBLICATIONS

R. Dubey, V. Banerjee, S. Hounsinou, G. Bloom, *Strong APA scheduling in a real-time operating system: work-in-progress*, International Conference on Embedded Software (EMSOFT), 2021. [DOI], [Talk], [Poster]

RECENT PROJECTS

Improving LLMs' performance on mathematical operations *Fall 2024*
- Enhanced GPT-2's performance on arithmetic tasks by developing custom computation modules and fine-tuning on specialized maths dataset, achieving **0.002 average error** after 10k training epochs.

Homography Projection for Surfaces in Art Paintings *Fall 2024*
- Built a **full-stack web app** (React, TypeScript, Docker, Streamlit) for distortion-free perspective correction of surfaces in paintings.

SKILLS

Programming Languages: C/C++14, Python, Java, SQL, JavaScript, TypeScript
Frameworks: PyTorch, CUDA, React, Spring Boot
Systems & Tools: Docker, Kubernetes, Terraform, Kafka, Linux, ElasticSearch