Tella Rajashekhar Reddy

Pre-Doctoral Research Fellow , Microsoft Research India

tella.rajashekhar@gmail.com
mr-rajashekhar.github.io
rajashekhar-reddy-tella
mr-rajashekhar

Education

2020 – 2024 **B.Tech. in Computer Science and Engineering,**
Indian Institute of Technology Dharwad
GPA: 9.5/10.0

Research Experience

July 2024 – present **Pre-Doctoral Research Fellow**, *Microsoft Research, India*
Mentors: Mentor: Debopam Bhattacherjee, Rohan Gandhi, Anjaly Parayil

- **AI Greenferencing**
  - Created AI Greenferencing to run AI workloads cost-effectively on renewable energy farms.
  - Built "Heron" cross-site router using power forecasts and spatial complementarity; hierarchical planners + deep power profiling (TP, GPU freq, RPS) to meet TTFT/TBT SLOs.
  - Won Microsoft's 2024 Hack for the Industry challenge among 20K+ projects and 75K+ participants.
- **LLM Congestion Control**
  - Built beLLMan, a system–LLM congestion-control interface that limits response length via load-aware prompt appending using TBT as the congestion signal.
  - Prototyped on vLLM (8× H100, Gemma-3 27B); built a output-length predictor (MAE 36, ~50 ms).
  - Achieved up to 8× lower E2E latency, ~25% less energy, and +19% throughput during congestion with minimal quality impact.

May 2023 – July 2023 **Mitacs Globalink Research Intern**, *Lakehead University*, Canada
Mentor: Dr. Shafiqul Hai

- **Hardware-Software Codesign of a Convolutional Neural Network**
  - Designed CNN for MNIST digits with 98% accuracy; optimized for 1M pixel images in 20ms.
  - Implemented block circulant matrix + 16-point FFT to reduce FPGA logical elements by 40%, maintaining 96.89% accuracy.

April 2022 – May 2024 **Undergraduate Research Assistant**, *Indian Institute of Technology Dharwad (IITDh)*, India
Affiliated to Future Generation Lab
Mentor: Dr. Koteswararao Kondepu

- Joined as a summer intern and continued for 4 semesters (R&D and B.Tech project); earned 1×AP (Exceptional Performance) and 3×AA grades.
- **vRAN Energy & Performance Profiling**
  - Assisted in deploying ORAN architectures (Monolithic, Dis-aggregated, and CUPS) and investigated their RAN and UE energy consumption using S-tui, Scaphandre and Kepler.
  - Profiled vRAN on next-generation processors to analyze the impact of CPU, cache, and memory on energy efficiency and throughput.
- **Open-Source FPGA toolchain and Compute Reservation Platform**
  - Developed an open-source FPGA toolchain for compilation, synthesis, and bitstream programming.
  - Built a platform for reservation of CPUs, GPUs, and FPGAs, optimising campus resource utilisation.
- **Edge-Assisted UAV Surveillance**
  - Engineered edge infrastructure for autonomous UAV navigation as part of the TiHAN project, leading the development of the object detection pipeline and autonomous navigation scripts.
  - Implemented a monitoring and alerting system using Prometheus, Grafana, and Kafka to manage UAV resource utilization.

Patents

P1. **TR. Reddy**, R. Gandhi, D. Bhattacherjee, "Memory Efficient Routing of Large Language Model Inference Requests", US Patent, Filed.

P2. **TR. Reddy**, D. Bhattacherjee, R. Gandhi, A. Parayil, C. Zhang, L. Yu *et al.*, "Cross-site Routing of Inference Workloads based on Predicted Power Availability", US Patent, Filed.

Journals

J1. S. Hai, **TR. Reddy**, "FPGA implementation of an Image Classifier Using Pipelined FFT Architecture", IEEE Embedded Systems Letters, 2024

Preprints

PP1. **TR. Reddy**, Palak, R. Gandhi, A. Parayil, C. Zhang, M. Shepperd, L. Yu, J. Mohan *et al.*, "AI Greenferencing: Routing AI Inferencing to Green Modular Data Centers with Heron", 2025

PP2. **TR. Reddy**, A. Deshmukh, K. Tandon, R. Gandhi, A. Parayil, D. Bhattacherjee, "BeLLMan: Controlling LLM Congestion", 2025

PP3. Palak, **TR. Reddy**, B. Kataria, R. Gandhi, K. Tandon, D. Bhattacherjee, VN. Padmanabhan, "Improving training time and GPU utilization in geo-distributed language model training", 2025

Publications

C1. **TR. Reddy**, U. Gupta, G. Venkateswarlu, V. R. Chintapalli *et al.*, "Resource Profiling for Virtualized Radio Access Networks", ANTS 2024.

C2. C. Centofanti, G. Venkateswarlu, J. Santos, **TR. Reddy** *et al.*, "An Energy Measurement Framework for 5G RAN Using USRP and Real-Time Monitoring", ANTS 2024.

C3. V. Gudepu, **TR. Reddy**, C. Centofanti, J. Santos, A. Marotta, K. Kondepu, "Demonstrating the Energy Consumption of Radio Access Networks in Container Clouds", NOMS 2024.

C4. **TR. Reddy**, S. Agarwal, K. Kondepu, "Exploiting Open Source Tools for FPGA Design Flow", COMSNETS 2024.

C5. N. Parekh, **TR. Reddy**, L. Malakalapalli, P. Tammana, K. Kondepu, "Real-Time UAV Resource Monitoring and Alerts with Automated Control Mechanism", COMSNETS 2024.

C6. V. Gudepu, B. Chirumamilla, **TR. Reddy**, A. Bhattacharyya, *et al.*, "EARNEST: Experimental Analysis of RAN Energy with Open-Source Software Tools", COMSNETS 2024.

C7. **TR. Reddy**, A. Marotta, P. Castoldi, L. Valcarenghi, K. Kondepu, "Enhancing UAV Systems via Task Offloading at the EDGE", ANTS 2023.

C8. Y. C. Makkena, **TR. Reddy**, N. Parekh, P. K. Saraf, H. Shukla *et al.*, "Experience: Implementation of Edge-Cloud for Autonomous Navigation Applications", COMSNETS 2023.

Teaching

Spring 2024 **CS-103: Data Structures and Algorithms**, *IITDh*
Instructors: Prof. Dileep A. D,Prof. Vandana Bharti

Autumn 2023 **CS-213: Software Systems Lab**, *IITDh*
Instructors: Prof. Koteswararao Kondepu

Spring 2023 **CS-102: Introduction to Programming in C & Python**, *IITDh*
Instructors: Prof. Ramchandra Phawade, Prof. Nikhil Hegde, Prof. Bharath B. N.

Relevant courses

Computer Science: Computer Programming, Data Base and Information Systems, Data Structures and Algorithms, Automata Theory, Design and Analysis of Algorithms, and Computer Architecture

Mathematics: Linear Algebra, Basic Calculus, Discrete Maths, and Probability & Random Processes

Technical Skills

Languages: C, C++, Python, Bash

Software Stacks: Container environments (Docker, Kubernetes), LLM serving (vLLM), Storage systems (Azure Blob Storage, MongoDB)

Tools & ML Frameworks: Jekyll, Git, Keras, TensorFlow

Achievements

- 1st Prize in the Hack for the Industry track at Microsoft Global Hackathon, 2024.
- Selected for the Google ML Bootcamp India.
- Awarded Conference Travel Grant and presented a research paper at ANTS 2023, Jaipur, India.
- Recipient of the Student Travel Grant Award at COMSNETS 2023.
- Achieved an exceptional AP grade in Linear Algebra , Research & Development and FPGA for Networks courses.
- 2nd Place in DevHack Hackathon organized by PARSEC, 2024.
- Served as Institute General Secretary, IIT Dharwad (2021–2022 academic year).
- Mentor at the Student Mentorship Program (SMP), IIT Dharwad.