Harsha Vardhan Reddy Kuncha

Stony Brook, NY | (623)-302-8447 | harshareddy.kuncha@gmail.com | linkedin.com/in/harshakuncha/

EDUCATION

Arizona State University — **Tempe, USA**
Master of Science in Computer Science | CGPA: 4/4 — Aug 2023 - May 2025
Coursework: Analysis of Algorithms | Operating Systems | Machine Learning | Distributed Systems

Manipal Institute of Technology — **Manipal, India**
Bachelor of Technology in Computer Science | CGPA: 8.1/10 — Aug 2018 - May 2023

SKILLS

Languages: C, C++, C#, Python, Go, Java, Bash, CUDA, JavaScript, TypeScript.
Technologies: Linux, PostgreSQL, MySQL, DynamoDB, AWS, EC2, S3, SQS, Lambda, Kubernetes, Docker, Tomcat, Postman, Swagger.
Frameworks: .NET, ReactJS, NodeJS, Flutter, Flask, Django, FastAPI, Angular, Node, Express, Spring Boot, RabbitMQ, Hibernate.
Certifications: AWS Cloud Practitioner, Oracle Cloud Infra Foundational(OCI), Big Data and Spark, Cisco Networking.

PROFESSIONAL EXPERIENCE

Software Engineer — May 2024 – Apr 2025
Growealth — Remote, Tempe,AZ.

- Spearheaded development of real-time **Java** dashboards using **Spring WebFlux**, **MVC (Thymeleaf)** and **D3.js**; enforced **PCI DSS**, and transformed tax reports into on-demand insights, driving a **22%** adoption uplift, worked on pipelines to cut scoring latency by **30%**.
- Built Java Spring Boot REST services with **Hibernate JPA**, Spring Batch for dispute-resolution to surface $3.2M in savings,integrating **Splunk** data logs and slashed **API latency by 40%** through Spring WebFlux and tested with **postman** and documented with **swagger**.
- Analyzed **300+** client portfolios via Java **Spring WebFlux** reactive streams to **Oracle DB**; implemented comprehensive unit tests (JUnit, Mockito) and **integration tests** to achieve **95% coverage** and involved in designing transaction-dispute patterns, fueled a client-personalized billing engine

Software Engineer — Jul 2022 – Jul 2023
Oracle — Hyderabad, India

- Architected and implemented a **React** + **TypeScript** front-end dashboard that visualized real-time analytics from backend event streams which were managed as a batch by Kafka, increasing user engagement by 30% and reduced page load times by 90%.
- Designed and built high performance RESTful APIs in **Java and Spring Boot**, deployed on **AWS Lambda** behind API Gateway leveraging asynchronous processing and aggregation pipelines to boost and achieved a reduced average request latency by 40%
- Developed **Node.js microservices** from legacy monolith architecture for order-processing workflows, containerized with **Docker** and orchestrated in Kubernetes on **AWS EKS**, handling 5 k TPS with 99.9% uptime which targets to serve 100M+ transactions monthly.
- Deployed an **AI-driven**, distributed system that tested **LLM outputs** in **RabbitMQ** event processing pipelines, improving data throughput by 25%; instrumented the platform with **Grafana** and **Splunk** for continuous monitoring and troubleshooting.
- Automated end-to-end **CI/CD** pipelines with GitLab CI and AWS Code Pipeline, integrating Terraform and CloudFormation to spin up S3 buckets, Lambda functions, and IAM roles cutting infrastructure provisioning time by 80%.

Project Intern — Jan 2022 – Jul 2022
Oracle — Hyderabad, India

- Designed and implemented web application layout with **Next.js**, **TypeScript, HTML and CSS**. that ingested WebSocket-driven event streams to display live system health metrics, reducing incident detection time by 40%.
- Wrote **Python** and automation scripts to batch-process and cleanse 50 GB+ of legacy data in Oracle cloud streams, loading it into **SQL** on RDS and improved unit and integration tests with pytest, jest and mockito maintaining 95% test coverage.
- Built serverless workflows with **AWS SQS** and **Lambda** (Node.js), processing user uploads asynchronously and improving resiliency.
- Worked closely and communicated with cross-functional teams in an agile environment to develop secure APIs, delivering **15+ tickets per sprint** with an average resolution time of **2 days**, ensuring timely delivery using Jira with **Scrum** methodology.

Software Engineer Intern — May 2021 – Nov 2021
Samsung — Hyderabad, India

- Engineered a **5G RAN base-station** simulator in NS-3 using **low-level C**, modeling radio propagation and handover events.
- Analyzed key performance indicators (**SINR**, throughput, handover success rate) across multiple **MIMO** configurations, uncovering a 15% degradation in link performance during high-mobility transitions

PROJECTS

Green AI | Python, Ollama, GPT+. — **Feb 2025 - Apr 2025**

- Engineered an end-to-end **Python** pipeline to analyze and optimize **prompt energy usage** for **Mistral-7B,** including data ingestion (5,000 prompts), feature extraction (token counts, punctuation density) and preprocessing with pandas and NumPy.

vLORA | LoRA, CUDA, GPU — **Jun 2024 - Dec 2024**

- Orchestrated the integration of Segmented Gather Matrix-Vector multiplication **CUDA** kernels of **Punica**, resulting in a remarkable **12x speed** for serving multiple LoRA models, **optimized GPU utilization** and reduced inference latency, enabling high-throughput.