Kalyani Vaidya | Software Engineer

989- 546- 9124 • vaidy1k@cmich.edu • LinkedIn• GitHub • Portfolio • Leetcode

EXPERIENCE

AI Research Assistant, Central Michigan University**,** MI, USA **Jan 2026 – Present**

- **Scaled system throughput to process 500+ concurrent tool-call requests** with sub-10ms parsing overhead by architecting highly concurrent, asynchronous GenAI inference middleware in **Python** to bridge third-party LLMs and MCP agents.
- **Reduced end-to-end round-trip inference latency by 40%** for LLM-driven agentic workflows by profiling bidirectional JSON-RPC payloads and directly resolving network-level I/O bottlenecks.
- **Enabled zero-blocking execution on the main Python event loop** by engineering a highly scalable server ecosystem that seamlessly integrates 10+ enterprise-grade infrastructure tools (including **Kubernetes** and AWS CloudWatch) into generative AI models.
- **Decreased data serialization overhead by 30%** to enable real-time processing of dynamic prompt arrays by optimizing the parsing, interception, and transmission of complex JSON-RPC payloads across distributed GenAI pipelines.
- **Ensured zero message loss and graceful error handling** during stress tests of 10,000+ sequential AI operations by designing robust, non-blocking observability layers and fault-tolerant routing systems for bidirectional server communication.

AI Software Engineer, Central Michigan University, MI, USA **Aug 2024 – Dec 2025**

- **Architected a distributed deep learning training framework** utilizing **Rust** and **C++**, optimizing Linux kernel schedulers and TCP networking stacks to achieve a 40% reduction in inter-node communication latency across a 50+ GPU cluster.
- **Engineered a high-performance, full-stack AI inference engine** backend using **Go** and **Python** (FastAPI/gRPC), successfully serving large-scale machine learning models at 10,000+ requests per second with sub-5ms latency.
- **Spearheaded end-to-end system profiling and debugging** for complex AI workloads using eBPF and OpenTelemetry, diagnosing root-cause TCP packet-drop bottlenecks at the edge and increasing overall system throughput by 35%.
- **Deployed and scaled state-of-the-art LLM and RAG pipelines** across distributed environments using Ray and Kubernetes, optimizing vector database querying algorithms to reduce AI inference time by 50% while maintaining 99.9% availability.
- **Proactively enhanced code quality and system reliability** by designing robust CI/CD pipelines (Docker, GitHub Actions) for ML model deployment, automating rigorous performance regression testing and reducing production rollout time by 60%.

Software Engineer – ML, APRG Technologies Pvt. Ltd, India **Aug 2022 - Nov 2023**

- **Reduced P99 inference latency by 35%** across 1M+ daily API requests by architecting high throughput, distributed Generative AI inference pipelines using FastAPI and **NVIDIA Triton Inference Server**.
- **Cut GPU VRAM requirements by 50%** and inference latency by 40% for enterprise-scale serving by productizing PyTorch deep learning models using **TensorRT** and INT8 quantization to maximize hardware-level execution efficiency.
- **Supported 5,000+ concurrent enterprise users** with fault-tolerant request routing by designing and deploying highly scalable distributed RESTful APIs on **Docker** and **Kubernetes (K8s)**.
- **Accelerated model iteration cycles by 3x** by engineering end-to-end distributed ML training workflows (PyTorch DDP), optimizing data ingestion pipelines to efficiently process and route multi-terabyte relational datasets.

TECHNICAL SKILLS

Languages: Python**,** Go, Rust, C++, SQL
Distributed Systems & Architecture: Kubernetes (K8s), Docker, Microservices, RESTful APIs, JSON-RPC, gRPC
System Optimization & OS: Linux Kernel Schedulers, TCP/IP Networking, Edge Computing, eBPF
Observability, Tracing & Debugging: Distributed Tracing, OpenTelemetry, AWS CloudWatch, Network I/O Profiling
AI/ML Infrastructure: NVIDIA Triton Inference Server, TensorRT, PyTorch Distributed Data Parallel (DDP), Model Quantization (INT8), CUDA

EDUCATION

Central Michigan University, Master of Science in Computer Science, MI (**GPA: 3.64 / 4.0**) Jan 2024 – Dec 2025

Savitribai Phule Pune University, Bachelor of Engineering in Information Technology, India (**GPA: 8.67 / 10.0**) Jun 2018 – May 2022

PROJECTS

CloudShop Lite: AI-Ops & Cloud-Native Microservices Platform GitHub

- **Increased API throughput by 3x** for distributed request routing by architecting a production-grade microservices ecosystem on AWS EKS, successfully migrating a legacy monolithic application to Python (FastAPI) and Nginx.
- **Decreased Mean Time To Recovery (MTTR) by 60%** for cluster failures by engineering a real-time AI-Ops automation engine that parses AWS CloudWatch logs and dynamically triggers self-healing Kubernetes remediations.

EcoBeanAI – Climate Yield Predictor (Published at IEEE AIBThings 2025) GitHub

- **Achieved high-precision predictive accuracies** of R^2 0.95 (quality) and 0.90 (yield) by engineering a scalable multi-target ML pipeline, integrating a PyTorch-wrapped XGBoost architecture with CTGAN synthetic data validated via rigorous K-S statistical testing.

PUBLICATION

- **Co-author, *EcoBeanAI: Predictive Modeling of Climate Effects on Coffee and Cocoa Yield*, IEEE AIBThings 2025.** Link