Aditya Prasath Ravilla Shridhar Prasad

Email: apr11@illinois.edu
Portfolio: Aditya's Engine Room
Mobile: +1 217-249-4900
Github: Aditya's Playground

Experience

- **StitchStudio** — Chicago, IL, Remote
  AI Engineering Intern — *May 2026 – Present*
  - **LLM Orchestration & Agentic Workflows**: Architecting scalable AI agents using **LangChain** and **Llama 3** for multi-step reasoning; implemented **dynamic prompt-routing** and state-machine logic to optimize token throughput and response precision.
  - **High-Performance Vector Retrieval**: Engineered **RAG (Retrieval-Augmented Generation)** pipelines leveraging **FAISS** for vector indexing, achieving **sub-100ms latency** across million-scale document embeddings for real-time semantic search.
- **Cognizant Technology Solutions (Client: Verizon)** — Tamil Nadu, India
  Software Engineer / Decisioning Consultant (Full-time) — *Aug 2023 – Jul 2025*
  - **High-Concurrency Distributed Systems**: Architected production microservices handling **8M+ daily requests**; maintained **99.95% availability** by implementing circuit breakers and automated horizontal scaling to handle high-traffic bursts.
  - **Performance & Latency Engineering**: Spearheaded a latency reduction initiative achieving a **42% improvement** in end-to-end response times by migrating bottlenecks to an **asynchronous Kafka-driven architecture** and optimizing SQL execution plans.
  - **Observability & Reliability**: Resolved critical memory leaks in distributed services using **Prometheus** and **Grafana**, reducing production incidents by **35%** while increasing test coverage by **25%** via automated suites.
- **SMZ & CO.** — Kuala Lumpur, Remote
  Backend Platform Intern — *Sep 2021 – Dec 2022*
  - **Scalable API & Schema Design**: Designed RESTful APIs and normalized **PostgreSQL** schemas to process **100K+ records/month**, reducing data reporting latency from minutes to **sub-second response times** through query optimization.

Projects

- **GPT-2 High-Performance Inference Engine (CUDA, C++, Nsight)**: Architected a functionally correct GPT-2 forward pass from scratch, implementing optimized CUDA kernels for **FlashAttention-2** and **KV-Caching** to eliminate redundant computations. Leveraged **Nsight Compute** to identify bottlenecks, implementing **Shared Memory Swizzling** and **Tiling** to resolve bank conflicts and maximize L2 cache hit rates on **NVIDIA A40** nodes.
- **ThinkerCUDA: High-Performance GPU Computing Library (CUDA, C++)**: Engineered custom CUDA kernels for 3D convolution and tiled matrix multiplication, resulting in a **6x throughput increase** over standard CPU implementations. Applied advanced memory-aware techniques, including shared memory tiling and memory coalescing, to maximize thread occupancy and minimize global memory latency.
- **Intelligent Audit Orchestrator (LangChain, Python, RAG)**: Architected an AI-driven task management system utilizing the BMAD-METHOD to automate specialized audit workflows. Implemented agentic prompt-routing strategies via LangChain to dynamically allocate resources based on task complexity, increasing throughput and accuracy.
- **Cloud-Native Auto-Scaling Platform (AWS, Kubernetes, Docker)**: Architected a microservices-based platform on AWS using EKS and EC2. Configured Application Load Balancers (ALB) and dynamic auto-scaling policies to maintain system performance during burst traffic, ensuring zero-downtime deployments via Blue-Green and Canary strategies.

Education

- **University of Illinois Urbana-Champaign** — Illinois, USA
  Master of Science in Computer Science; GPA: 3.91 — *Aug 2025 - Pursuing*
  Key Coursework: *Applied Parallel Programming, Systems for GenAI, Applied Machine Learning, Data Management, Cloud Computing, LLMs*
- **SRM Institute of Science and Technology** — Tamil Nadu, India
  B.Tech in Computer Science & Business Systems; GPA: 3.97 — *Jun 2019 - Jul 2023*
  Key Coursework: *Data Structures and Algorithms, Formal Language and Automata Theory, OOP, DBMS, Compiler Design, Computer Networks*

Technical Skills

- **Programming**: Python (NumPy, Pandas), C++, Java, JavaScript, Bash, SQL, Git.
- **Performance & Systems**: High-Throughput Processing (8M+ requests/day), Latency Reduction (42% improvement), GPU Acceleration (CUDA), Query Optimization.
- **Cloud & Infrastructure**: Cloud-Native Architecture (AWS), Docker & Kubernetes, Microservices, RESTful API Design, Scalable Backend Design, Prometheus, Grafana.
- **AI & Data Engineering**: LLMs (Llama 3, DeepSeek), LangChain, RAG-based Analytics, BMAD-METHOD, Spark, HDFS, ETL Pipelines, FAISS.

Research & Awards

- **PACT: Pruned Agent Call Throughput (UIUC CS 598)**: Architected a decoupling strategy using a fine-tuned **Speculator (Phi-3-mini)** to dynamically truncate LLM over-deliberation; implemented **DPO preference alignment** to optimize accuracy-latency trade-offs, aiming to reduce agentic response times from minutes to seconds.
- **Intelligent Surveillance over 5G Edge Networks**: Conference publication on optimizing latency and bandwidth for real-time AI inference pipelines; evaluated compute offloading strategies across edge and centralized nodes.
- **Outstanding Contribution Award**: Dean's recognition for exceptional academic and institutional contributions throughout the Bachelor's degree program (2019–2023).