Yashwanth Reddy Dasari

312-646-8975 | yashwanth2632@gmail.com | Portfolio | LinkedIn | Github

Software Engineer specializing in backend systems, distributed data pipelines, and AI-driven applications. Experienced in building scalable microservices, semantic search platforms, and multi-agent LLM systems, shipping production features end-to-end.

Education

University of Illinois Chicago (*Master of Science in Computer Science) CGPA - 3.9/4* Aug 2023 – May 2025

Vellore Institute of Technology (VIT) (*Bachelor of Science in Computer Science) CGPA - 3.7/4* Jul 2019 – May 2023

Relevant Coursework: *Operating Systems, Database Systems, Distributed Systems, Computer Networks, Compiler design, Design and Analysis of Algorithms, Machine Learning, Optimization Techniques, Computer Architecture, Neural Networks.*

Technical Skills

- **Programming Languages:** Python, Java, Go, C++, SQL, TypeScript, JavaScript, CUDA
- **Backend & Systems:** Spring Boot, FastAPI, Django, Microservices, Kafka, gRPC, Redis, Elasticsearch, RocksDB, Raft Consensus, PostgreSQL, MongoDB, Cassandra, DynamoDB.
- **AI & Data Systems:** PyTorch, vLLM, DeepSpeed, Vector Search (HNSW), Embedding Pipelines, CUDA Optimization
- **Cloud & DevOps:** Docker, Kubernetes, AWS, GCP, Azure, CI/CD Pipelines, Prometheus, Grafana, NGINX, Git.
- **Frontend:** React.js, Next.js, D3.js, Node.js, WebSockets, Tailwind CSS.
- **Testing & Quality:** Unit Testing, Integration Testing, E2E Testing, PyTest, JUnit, Contract Testing

Work Experience

UICenter - University of Illinois, *Software Engineer, Chicago, IL* Jan 2024 – Present

- Engineered a high-throughput event ingestion service (Spring Boot + Kafka) using idempotent consumers and DLQs, ensuring zero-loss processing for 5M+ daily research events.
- Designed and implemented a semantic search and retrieval pipeline using Elasticsearch and vector embeddings, achieving ∼150ms median latency and replacing multi-hour manual lookups for 1,000+ researchers.
- Accelerated compute workflows 3× (6h → 2h) by offloading matrix multiplications to A100 GPUs with custom CUDA kernels optimized via shared-memory tiling.
- Developed cloud-native resilient REST/gRPC microservices using Resilience4j circuit breakers and Redis caching to prevent cascading failures during peak ingestion loads.
- Developed internal npm libraries and shared UI/service utilities to standardize workflows across 10+ research tools, reducing duplicated logic and accelerating feature delivery.

ImpacterAI Inc (Early Stage Startup), *Software Engineer Intern, San Francisco, CA* Aug 2025 – Nov 2025

- Developed a FastAPI backend for multi-agent LLM orchestration, implementing async task routing, shared memory layer, and self-refining 'Judge' loops that reduced redundant API calls by ∼30% and increased workflow throughput by ∼5×.
- Migrated production inference to vLLM, optimizing KV cache utilization via PagedAttention, improving GPU utilization by ∼40% while maintaining low Time-To-First-Token (TTFT) for enterprise clients.
- Built a real-time observability console (React + Next.js + WebSockets) visualizing 80k+/min inference logs to support live debugging and reduce MTTR by 60%.

SimplyFI Innovations, *Software Engineer 1, India* Jan 2023 – Aug 2023

- Built backend services for a trade-finance platform processing $25M+ daily volume, developing Django REST APIs and NLP-driven automation pipelines achieving ∼98% verification accuracy.
- Refactored and decomposed a legacy monolith into 16+ microservices (Docker + Kubernetes) with Kafka event workflows, reducing system latency by ∼35% and improving UI responsiveness.
- Implemented comprehensive test suites (unit, integration, E2E) across 25+ microservices using PyTest and JUnit, achieving 85%+ code coverage and improving deployment stability.
- Implemented OAuth2/RBAC workflows to enforce Zero Trust security policies, improving traceability and reducing incident investigation time from 45 → 15 minutes.

Projects

Distributed Consensus KV Store | Go, gRPC, RocksDB, Raft

- Implemented a Raft-based distributed KV store with leader election, log replication, and linearizable consistency, ensuring system availability during network partitions across a 5-node cluster.
- Designed a RocksDB-backed storage layer using LSM-tree batching, compaction, and snapshot recovery, utilizing Write-Ahead Logs (WAL) to guarantee durability under high write loads.
- Integrated Prometheus/Grafana and a React dashboard to visualize replication lag, commit index, and cluster health in realtime.

High-Performance Vector Search Engine | C++, FastAPI, Python, SIMD

- Built an HNSW-based ANN engine via AVX2 SIMD optimizations for top-k search, enabling low-latency embedding retrieval.
- Implemented Product Quantization (PQ) achieving 4× memory reduction while preserving high recall on 10M+ vectors.
- Exposed C++ core via Pybind11 and FastAPI, enabling concurrent, low-latency serving for scalable, real-time ML pipelines.