Keshav Jha

SOFTWARE ENGINEER

☎ +919599015933 | ✉ keshavsde@gmail.com | therealsaitama | CodesPasta

Education

Delhi Technical Campus (GGSIPU) *Greater Noida, India*

B.Tech. in Computer Science — *Expected Nov 2026* *Nov 2022 – Present*

- **Relevant coursework:** Data Structures & Algorithms, Operating Systems, Computer Networks (TCP/IP), Distributed Systems, Database Systems, Information Retrieval, Machine Learning.

Experience

She&Soul (Startup) *New Delhi, India*

Co-founder & Lead Backend Engineer *2024 – Present*

- Architected the platform backend using **FastAPI**, **PostgreSQL**, and **Redis**; designed REST APIs, JWT auth, RBAC, and a multi-tenant data model.
- Implemented payment, order, notification workflows (**FCM**/SMS/email) with rate limiting and request tracing; added structured logging and retries.
- Containerized services with **Docker** and **Nginx** reverse proxy; set up **GitHub Actions** CI and zero-downtime migrations.

Projects

spot-http — Non-Blocking HTTP Server

Systems • C++/Linux *Oct 2025 – Present*

- Built an **epoll**-based HTTP/1.1 server with zero-copy **`sendfile()`**, connection pooling/keep-alive, static file cache, and pluggable router/middleware.
- Added graceful shutdown and back-pressure; wrote k6/Locust load scripts and perf profiling to validate throughput and tail latency under burst traffic.
- **Results (fill in):** sustained *N* RPS on 4-vCPU VM with P95 ≤ *X* ms; zero errors at *M* concurrent conns.

event-fanout — WebSocket Fan-out Service

Distributed Systems • Go *Oct 2025 – Present*

- Gateway in **Go** using goroutines/channels with room-based broadcast, heartbeats, and back-pressure-aware writers.
- Horizontal fan-out via **Redis Pub/Sub** (swappable to NATS); sharded channel registry; graceful shutdown and connection drain.
- **Results (fill in):** *K* concurrent clients; *M* msgs/s broadcast; P95 latency ≤ *T* ms in load tests.

mini-search — BM25 Inverted Index + FastAPI

Information Retrieval • Python *Oct 2025 – Present*

- Tokenized text and built an on-disk **inverted index**; implemented **BM25** ranking, snippet highlighting, and a **/search** API.
- Benchmarked warm vs. cold cache; added index-rebuild tooling; shipped a Dockerized deployment for reproducible runs.
- **Results (fill in):** indexed *K* docs in <*T* s; served *Q* QPS with warm-cache P95 ≤ *Y* ms.

ml-serve-ab — ONNX Inference with A/B + Canary

ML Systems • Python *Oct 2025 – Present*

- FastAPI microservice using **ONNX Runtime** (ResNet50) with **80/20** A/B routing and canary flags; Redis cache; **/healthz** and **/metrics**.
- SRE basics: 99.9% SLO template, Prometheus/Grafana alerts, graceful shutdown, retries with jitter, and idempotency keys.
- **Latency/throughput (fill in):** CPU P95 ≤ *A* ms, cache hit ≥ *B*%, GPU P95 ≤ *C* ms; SLO met over *H* hrs.

Skills

Languages	Python, C++, Go, SQL
Systems & Networking	Linux/UNIX, TCP/IP, Sockets (*epoll*), WebSockets, HTTP/REST
Frameworks & Datastores	FastAPI, Flask, PostgreSQL, Redis, Docker, Nginx
ML & IR	PyTorch, ONNX Runtime, scikit-learn, NumPy, BM25/TF-IDF
DevOps & Tooling	Git, GitHub Actions, Bash/Zsh, Prometheus, Grafana, K6/Locust

Achievements

2025	Solved **2000+** DSA problems across LC/CF/GFG/HackerRank
2024	LeetCode **Knight** (Top **3%**, Max **1938**)
2024	CodeChef **5 star** (Max **2077**); India Rank **540**
2024	GeeksforGeeks Max **2180** (Global #**41**)
2024	Codeforces Round 984 Global #**327**/50k
2024	CodeChef Starters 163 India #**27** (Global #34)
2024	GFG Weekly 180 Global #**19**