

ALEXEI KHACHATRIAN

AI Infrastructure & Backend Platform Engineer • 10+ Years • Production Systems at Scale

Yerevan, Armenia • alexey.khachatryan@gmail.com • +374 98 888 142

khachatryan.dev • github.com/alkhachatryan • linkedin.com/in/alexei-khachatrian

SUMMARY

Senior backend engineer building the infrastructure that AI runs on — high-throughput inference paths, data pipelines, vector storage, multi-tenant API platforms, and event-driven systems. 10+ years operating production backends at scale, with current systems handling ~600 RPS (architected toward 1,000) and prior work on a ~2 PB PostgreSQL analytics platform. Active PhD researcher building Python automation pipelines for protein analysis on an HPC supercomputer — production ML/AI workloads, not theory. Led distributed engineering teams across five countries. Deep AWS / GCP / Azure background. Fluent across PHP, Python, Node.js, C/C++.

AI INFRASTRUCTURE & PLATFORM CAPABILITIES

- **High-throughput backend APIs:** operating systems at ~600 RPS today, architecting to 1,000 — the same patterns used in LLM inference serving, model gateway routing, and token-streaming endpoints.
- **Large-scale data infrastructure:** petabyte-class PostgreSQL (~2 PB) at Wildbox; query optimization, ingestion pipelines, multi-million-row contact datasets at iNexus — directly applicable to embedding stores, training data pipelines, and RAG retrieval.
- **Production ML / scientific compute (active):** PhD research building Python pipelines for protein analysis on the Aznavour supercomputer — protein-bank data aggregation, an AI-driven recommendation system based on user behavior, and a drug-interaction simulation pipeline running every candidate molecule against the full human proteome.
- **Event-driven & async systems:** queue infrastructure (SQS, Supervisor, Laravel queues, Redis pub/sub), Server-Sent Events for live data streaming, Socket.IO — the same primitives used in agent orchestration, streaming inference, and async LLM workloads.
- **Multi-tenant SaaS & billing platforms:** built end-to-end with Stripe Cashier, tiered plan enforcement, ~40-permission RBAC, Keycloak SSO/JWT — directly maps to AI-API billing, usage metering, and per-tenant model access control.
- **Cloud-native operations:** production workloads across AWS, GCP, and Azure in parallel; Docker, CI/CD, CloudFront, queue systems, observability.
- **Search & retrieval:** Elasticsearch + Kibana in production across multiple platforms — foundational for RAG, hybrid search, and embedding-augmented retrieval pipelines.

EXPERIENCE

Senior Software Development Engineer — NGA — Next Generation Advanced *Dec 2023 — Present*

- Building automation software for emergency-services platforms deployed internationally (primarily US).
- Operating production backend at ~600 requests/second; architecting the path to 1,000 RPS. Reliability and latency are non-negotiable — the same constraints AI inference platforms face under load.

PhD Researcher — Bioinformatics Lab Collaboration — European University of Armenia /

Bioinformatics Group

Mar 2025 — Present

- Applied research at the intersection of computer science and bioinformatics, with production-grade Python pipelines running on the Aznavour HPC supercomputer.

- Built a protein-bank data aggregator that pulls from RCSB PDB and other major sources into a unified internal dataset.
- Built an AI-driven recommendation system that infers researcher interests from browsing behavior across the protein catalogue and surfaces related entries — production behavioral signal → recommendation pipeline.
- Built a drug-design automation pipeline that takes a candidate molecule from a pharmaceutical lab and simulates its interaction against every human protein in multiple binding positions to predict therapeutic or toxic effect.

Python Developer — Wildbox.ru

Mar 2023 — Dec 2023

- Built analytics features on a ~2 PB PostgreSQL dataset for a Wildberries-marketplace analytics platform.
- Query performance optimization, ingestion pipelines, and feature work under heavy data volume — directly applicable to large-scale embedding stores and training-data warehouses.

Lead Software Development Engineer — Akvelon, Inc.

Mar 2022 — Mar 2023

- Led a developer team building an information-security SaaS B2B product.
- Architected and operated infrastructure across Azure, GCP, and AWS in parallel.
- Ran one-week sprints, stand-ups with client and team, code reviews, CI/CD pipelines.

Senior Software Development Engineer — Zamigos Consulting Group Aug 2020 — Mar 2022

- Delivered backend systems across financial, messaging, and accounting verticals.
- Worked across multiple stacks and teams; collaborated closely with DevOps, frontend, and QA.

Lead Software Development Engineer — iNexus

Oct 2017 — Mar 2020

- Built Virtual Publicist from scratch as lead engineer — a music-industry PR platform on Laravel + PostgreSQL + Redis + Elasticsearch, scaled to a multi-million-row contact dataset before public release.
- Recruited and led a team of two frontend engineers, one backend engineer, and a QA over three years.
- Owned security: penetration testing and exploit testing across our infrastructure and client sites.

Freelance Backend Lead (side work) — Self-employed

Apr 2017 — Aug 2020

- Led a distributed backend team across Spain, Russia, Ukraine, Armenia, and Israel for an Israeli-American e-commerce project.

SELECTED PROJECTS

JobsRadar — *Founder & sole engineer — production*

- Global remote-job aggregation platform. Built solo and operated in production. Ingests listings from multiple job-board sources into Laravel + PostgreSQL + Redis with Python pipelines; cron-based sync every few hours; CloudFront + AWS infrastructure; SEO-driven indexing. jobs-radar.com

ReserveX — *Co-founder & lead engineer — B2B booking SaaS (discontinued)*

- Multi-tenant white-label SaaS with per-company subdomains. Laravel 10 + PHP 8.2 API, JWT auth, Stripe Cashier billing with tiered plan limits enforced via middleware, ~40-permission RBAC, calendar service with conflict resolution, Elasticsearch-powered discovery, AWS + CodeDeploy + Sentry + Docker.

Activeo — *Founder & sole engineer — Gym management SaaS (discontinued)*

- Multi-tenant SaaS for gym operators. Laravel backend, React dashboard, Stripe billing with subscription enforcement, Keycloak SSO/JWT auth, SSR public sites, multilingual content with fallback, tenant isolation, AWS deployment.

MeshLink — *ESP-NOW sensor mesh + FastAPI dashboard (open-source)*

- Wireless sensor mesh designed end-to-end. ESP32 nodes broadcast environmental data over ESP-NOW with 16-bit network-token validation, max-hops limit, 64-entry ring-cache deduplication. FastAPI server with 3-hour rolling time-series per node, live Server-Sent Events to a browser dashboard.

DTLP — Data Transmission Laser Protocol — *Custom optical protocol on dual ESP32 (open-source)*

- Bidirectional laser-based communication protocol. Eight parallel laser channels yield ~8 kbit/s by transmitting one byte in parallel per millisecond. ESP-IDF + FreeRTOS with task pinning, custom handshake, bit-level synchronization, optional XOR encryption and LZ4 compression.

SKILLS

Languages: Python (Django, Flask, FastAPI), PHP (Laravel), JavaScript / Node.js, C / C++, Bash, SQL

Backend & Platform: REST / JSON APIs, microservices, multi-tenant SaaS, event-driven systems, Server-Sent Events, Socket.IO, Stripe / Cashier, Keycloak SSO / OAuth / JWT, Selenium

Data & Search: PostgreSQL (incl. ~2 PB scale), MySQL, MongoDB, DynamoDB, Redis, Elasticsearch + Kibana

AI / ML adjacent: Vector-store-ready Postgres (pgvector-compatible patterns), embedding-friendly retrieval (Elasticsearch hybrid search), Python ML/HPC pipelines on supercomputer, recommendation systems, behavioral signal ingestion, RAG-pattern infrastructure (ingestion → store → retrieval → API)

Cloud & Infra: AWS (EC2, RDS, S3, CloudFront, Route 53, Cognito, Elastic Beanstalk, CodePipeline/Deploy/Build, CloudWatch, SES, SNS, SQS, IAM, DynamoDB), GCP, Azure, Docker, Nginx, Supervisor, CI/CD

Scale & Performance: High-throughput production (~600 RPS today, designed toward 1,000), petabyte-scale databases, multi-million-row datasets

Embedded / IoT: ESP32, ESP-IDF, FreeRTOS, ESP-NOW, ESP-MESH, LoRa / LoRaWAN, custom protocol design

Leadership: Distributed team leadership, hiring, estimation, risk planning, technical design

Operations & Security: Ubuntu Server, CentOS, Kali Linux, penetration & exploit testing, CI/CD, GitLab / GitHub

EDUCATION

PhD in Computer Science (in progress) — European University of Armenia, 2024 — 2029

Research focus: software automation. Applied work in bioinformatics — protein-bank data aggregation, AI-driven research recommendation, and drug-interaction simulation on the Aznavour supercomputer.

MSc in Computer Science — European University of Armenia, 2021 — 2023

BSc in Computer Science — European University of Armenia, 2011 — 2017

CERTIFICATIONS & OPEN SOURCE

- Laravel Certified Developer — Laravel.com (2021)
- TestDome PHP — Gold (Top 10% Worldwide, 2018)
- Active open-source contributor since 2017 — author and maintainer across the Electr0Hub, data-transmission-laser-protocol, and krypt0ra GitHub organizations.

LANGUAGES

English (professional), Russian (native), Armenian (native).