SERGEY TOVPEKO

Platform & AI Infrastructure Engineer

Kubernetes · MLOps · Go/Python · High-Load Inference · Data Platforms

Email: aqu.de@yandex.ru **Telegram:** @rustnomicon **Social:** tovpekosergey.tech github.com/spgsroot

PROFILE

Platform / SRE engineer with deep expertise in automation, observability, and infrastructure for ML workloads. Specialized in high-load local model inference (vLLM/Triton), GPU utilization optimization, and event-driven RAG systems. Product-minded: I do not only configure infrastructure, but also build internal automation tools, CLI utilities, and lightweight client interfaces for monitoring and operating systems end-to-end.

TECHNICAL SKILLS

AI Infrastructure & MLOps: vLLM, Ollama, CUDA optimization, NVIDIA GPU Operator, RAG pipelines, pgvector, Qdrant, n8n.

Platform & DevOps: Kubernetes (k3s), Docker, GitLab CI, Automated Delivery Pipelines, Ansible, Linux, eBPF.

Backend & Observability: Python (FastAPI), Go, C#, PostgreSQL, ClickHouse, Apache Airflow, Prometheus, Grafana, VictoriaMetrics, ELK Stack, MinIO/S3.

Internal Tooling & Client Interfaces: Kotlin (Jetpack Compose), Flutter (Dart), REST/WebSocket clients, CLI utilities.

PROFESSIONAL EXPERIENCE

AO RTI | Platform / DevOps Engineer

Apr 2025 - Apr 2026

- Designed and deployed high-load LLM inference (Qwen) using vLLM and Ollama, optimized GPU utilization with CUDA, and reached 150+ tokens/sec on a single GPU.
- Implemented a scalable RAG pipeline with vector search (pgvector, Qdrant) and event-driven orchestration based on n8n and Supabase.
- Deployed and configured a ClickHouse cluster from scratch for infrastructure metrics collection.
- Automated ML model delivery into Kubernetes (k3s) and configured GPU metrics monitoring with Prometheus and Grafana.
- Built internal cross-platform client applications for IoT telemetry monitoring and remote diagnostics of microcomputers.

Sber | DevOps / Automation Engineer

Feb 2024 - Apr 2025

- Developed load-testing tooling: custom CLI wrappers and traffic generators in Go/Python.
- Improved system stability under load with Chaos Engineering and stress tests (k6, wrk).
- Created isolated Docker-based testing environments with emulation of heavy external services (Spark/Hadoop).
- Reworked observability by adding business metrics and long-term storage in VictoriaMetrics and ClickHouse.

AO "Rosseti Tsifra" | DevOps / Data Platform Engineer

Jul 2023 - Feb 2024

- Designed and operated a fault-tolerant process automation platform based on Apache Airflow and optimized DAG structure.
- Implemented centralized logging and tracing on ELK Stack, reducing mean time to recovery (MTTR).
- Containerized system components and standardized dev/prod environments with Docker.
- Administered distributed data stores (PostgreSQL, MariaDB, MinIO/S3) and tuned their performance.

Franklins Burger | Fullstack Developer / Automation Engineer

Dec 2022 - Apr 2024 · 1 year 5 months

- Built automated software update delivery infrastructure for endpoints in Yandex Cloud from scratch using Docker and Ansible.
- Designed and implemented a distributed data bus for centralized point configuration and advertising management.
- Developed server-side POS integration modules (IIKO) in C#/.NET and Python (FastAPI).
- Designed and launched internal web tools for business automation, CRM systems, and HR bots.

SELECTED PROJECTS

High-Performance ML Inference Serving: Local LLM deployment stand for Qwen/Llama based on vLLM/Ollama with autoscaling, CUDA monitoring, and production-like containerization. vLLM, Docker, CUDA, Prometheus, Grafana, Python.

AI-Driven Enterprise Automation Core: Agent workflow orchestration stack integrated with vector databases for contextual search, RAG, and internal knowledge automation. n8n, pgvector, Qdrant, FastAPI, Supabase.

Multi-Platform CI/CD Automation Framework: Universal software delivery automation template with containerization, artifact signing, release notes, and crash reporting integration. GitLab CI, Fastlane, Docker, Ansible, Gradle.

EDUCATION

Informatics and Computer Science (09.03.01)
Bauman Moscow State Technical University (Expected 2025)
Incomplete Higher Education

Information Systems and Programming (09.02.07)
MIREA – Russian Technological University (2024)
Associate / Special Secondary Degree

LANGUAGES

- **Russian:** Native
- **English:** B1 (Intermediate Professional)