

Andrey Krisanov

Moscow, Russia | develop4net@gmail.com | github.com/akrisanov | linkedin.com/in/akrisanov | akrisanov.com

Staff Software Engineer | LLM Inference Platforms & AI Infrastructure

PROFILE

Staff software engineer specializing in production LLM inference platforms, AI infrastructure, and distributed systems. At Severstal, I lead the architecture and hands-on technical evolution of DaVinci, a shared model-serving platform running self-hosted open-weight models on Kubernetes, vLLM, and 24 NVIDIA H200 GPUs.

My work covers inference control planes, model-serving performance and reliability, observability, GPU capacity planning, controlled model rollouts and rollbacks, and multi-data-center resilience. Previously, I built and scaled backend and platform systems across SaaS, fintech, data privacy, payments, and high-traffic consumer products.

SELECTED IMPACT

- Lead the architecture and hands-on technical evolution of Severstal's shared LLM inference platform, currently running on 24 NVIDIA H200 GPUs; designed its target topology, capacity allocation, and resilience model for expansion to 48 H200 and eight H100 GPUs across two data centers.
- Modernized a \$3M+ ARR SaaS platform, improving critical backend paths by 2–10x, achieving 99.998% availability, and reducing mean time to recovery by 50%.
- Designed and launched X5 Media, a platform that reached more than 20 million monthly active users within its first year.

WORK EXPERIENCE

Staff Software Engineer

Jan 2026 — Present

Severstal | DaVinci GenAI Platform

Moscow, Russia

- Lead architecture and technical direction for DaVinci, a shared GenAI platform serving enterprise AI products, assistants, and coding agents.
- Own production inference for self-hosted open-weight models on Kubernetes, vLLM, and 24 NVIDIA H200 GPUs, including model placement, parallelism, continuous batching, KV-cache allocation, context limits, and request limits.
- Design and develop the AI gateway and inference control plane for streaming, model routing, aliases, quotas, rate limits, admission control, fallbacks, evaluation, and controlled rollouts and rollbacks.
- Benchmark and tune serving configurations against production traffic, balancing time to first token, token throughput, concurrency, context length, and GPU memory.
- Designed capacity allocation, degradation policies, fail-over, and recovery for a planned expansion to 48 H200 and eight H100 GPUs across two data centers.
- Build SLO-driven operations with Prometheus, Grafana, GPU telemetry, alerting, load testing, incident analysis, and capacity forecasting.

Chief Technology Officer

Jan 2024 — Dec 2025

PlanFact | B2B Accounting SaaS

Moscow, Russia

- Served as the hands-on technical lead for modernization of a \$3M+ ARR SaaS product in a 100-person company while leading a 25-person technology organization.
- Modernized the core .NET platform through a domain-oriented modular architecture, replacing legacy dependencies without disrupting business-critical services.
- Improved critical backend paths by 2–10x and scaled read-heavy workloads through targeted refactoring, SQL and database tuning, strongly consistent replication, and horizontal read scaling.
- Introduced service-level observability and incident-management practices, achieving 99.998% availability and reducing mean time to recovery by 50%.

Senior Software Engineer, Backend & Platform

Stattice.ai | Data Privacy Platform

Jun 2021 — Dec 2023

Berlin, Germany

- Built Python, FastAPI, and Celery backend services and Kubernetes platform workflows on GCP for a cloud PII-protection product and privacy-preserving machine-learning workloads.
- Modernized SDK architecture and cross-platform x64/ARM build pipelines; introduced type checking, static analysis, CI improvements, and security scanning while maintaining 99.99% availability.

Engineering Lead, Backend & Platform

X5 | Retail Technology

Jun 2019 — Jun 2021

Moscow, Russia

- Designed and launched X5 Media, a Python, FastAPI, and Next.js platform that reached more than 20 million monthly active users within its first year.
- Owned backend and platform architecture during the transition from outsourced to in-house engineering; built the first cross-functional teams and halved time-to-market for existing products.

Earlier Engineering Experience

Fintech, SaaS, and Startups

2012 — 2019

Germany, Russia, and Remote

- **Independent consultant:** Migrated a Frankfurt fintech from on-premises infrastructure to AWS, cutting infrastructure costs by 50%; designed cloud architectures for two early-stage startups.
- **Clark:** Built a secure server-to-server integration with N26 that enabled in-app insurance signing and created a new partnership revenue stream; modernized and hardened a Ruby on Rails platform to improve security, performance, maintainability, and support for white-label products.
- **Ridewithlocal:** Owned backend engineering, AWS infrastructure, and production reliability for the initial product launch, maintaining 99.99% uptime; delivered the MVP on schedule and helped the company secure a \$300K pre-seed round.
- **Payler:** Designed and launched a PCI DSS-certified payment gateway, helping transform the company from a services business into a product-focused fintech platform; scaled engineering from one developer to a team of approximately 30.

SKILLS

- **LLM Inference & Serving:** vLLM, distributed model serving, continuous batching, KV-cache management, tensor and data parallelism, latency and throughput benchmarking, GPU capacity planning, NVIDIA A100/H100/H200
- **Inference Platforms:** AI gateways, model routing, quotas, rate limits, admission control, fallbacks, model life-cycle, controlled rollouts and rollbacks, multi-tenancy
- **Distributed Systems & Reliability:** Kubernetes, multi-data-center resilience, SLOs, capacity forecasting, load testing, Prometheus, Grafana, DCGM
- **Programming & Data:** Python, FastAPI, Go, C# (.NET), TypeScript, PostgreSQL, Redis, ClickHouse
- **Infrastructure:** Linux, Docker, Terraform, Ansible, AWS, GCP

EDUCATION

Bryansk State University

Master's in Applied Mathematics & Information Science

Bryansk, Russia

2006 — 2011