Andrey Krisanov

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

Staff Software Engineer, AI Infrastructure & LLM Inference

Profile

Staff Software Engineer specializing in production LLM inference, AI/ML infrastructure, and distributed systems. At Severstal, I lead the architecture and evolution of DaVinci, a shared inference platform supporting enterprise AI products, coding agents, and agentic workflows.

My work covers GPU-backed model serving, traffic management, performance and reliability engineering, observability, model lifecycle, capacity planning, and multi-data-center resilience.

14+ years of experience building backend, cloud, distributed, and SaaS systems, combining hands-on engineering with cross-team architecture and technical direction.

Selected Career Highlights

- Lead the architecture of Severstal's shared AI platform and inference foundation for enterprise AI products, coding agents, and agentic workflows, currently running on 24 NVIDIA H200 GPUs and expanding to 48 NVIDIA H200 and 8 H100 GPUs across two data centers.
- Modernized a $3M+ ARR SaaS platform, improving critical backend paths by 2–10x, reaching 99.998% availability, and reducing incident-resolution time by 2x.
- Designed and launched X5 Media, a Python-based content platform that grew to more than 20 million monthly active users within its first year.
- Designed and launched a PCI DSS-certified payment gateway and helped scale the engineering organization from one developer to approximately 30.

Work Experience

Staff Software Engineer Jan 2026 — Present
Severstal | DaVinci GenAI Platform *Moscow, Russia*

- Lead the architecture and technical direction of Severstal's shared GenAI platform, supporting enterprise AI products, assistants, coding agents, agentic workflows, and controlled tool execution.
- Own the production LLM inference architecture running self-hosted models on Kubernetes, vLLM, and 24 NVIDIA H200 GPUs; lead the expansion to 48 NVIDIA H200 and 8 H100 GPUs across two data centers.
- Design multi-data-center traffic distribution, capacity allocation, degraded operation, failover, and recovery to keep AI services available during infrastructure failures.
- Improve the performance and stability of DeepSeek, Kimi, and Qwen serving by analyzing production traffic and tuning batching, parallelism, queueing, KV-cache usage, and GPU memory allocation.
- Design and develop the AI gateway and model-delivery layer, covering model routing, aliases, quotas, rate limiting, fallbacks, evaluation, safe rollout, and rollback across self-hosted and external models.
- Establish SLO-driven operations through Prometheus and Grafana observability, alerting, load testing, incident analysis, and GPU capacity forecasting.

Chief Technology Officer (Hands-on Principal Engineer) Jan 2024 — Dec 2025
PlanFact | B2B Accounting SaaS *Moscow, Russia*

- Led technology strategy and hands-on platform modernization for a $3M+ ARR SaaS product in a 100-person company, covering backend architecture, reliability, cloud infrastructure, and a 25-person IT organization.
- Modernized the core .NET platform from .NET 6 toward .NET 9, replaced legacy dependencies, improved security, and kept business-critical services stable throughout the migration.
- Designed and introduced a domain-oriented modular architecture that replaced overloaded layered patterns and became the standard for new backend development.
- Improved critical backend paths by 2–10x and scaled read-heavy workloads through targeted refactoring, SQL optimization, database tuning, strongly consistent replication, and horizontal read scaling.
- Established service-level observability and incident-management practices, reaching 99.998% uptime and reducing incident-resolution time by 2x.
- Introduced Terraform and Ansible and raised engineering standards through architecture reviews, RFCs, code-review practices, mentoring, and hiring.

Senior Software Engineer, Backend & Platform Jun 2021 — Dec 2023
Statice.ai | Data Privacy Platform *Berlin, Germany*

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

Engineering Lead, Backend & Platform Jun 2019 — Jun 2021
X5 | Retail Technology *Moscow, Russia*

- Designed and launched X5 Media, a Python, FastAPI, and Next.js platform that grew to more than 20 million monthly active users within one year.
- Owned backend and platform architecture across consumer, B2B, and internal products during the transition from outsourced to in-house engineering.
- Built the first cross-functional in-house engineering teams and improved delivery processes, reducing time-to-market for existing products by 2x.

Earlier Engineering & Leadership Experience 2012 — 2019
Fintech, SaaS, and Startups *Germany, Russia, and Remote*

- **Independent consultant:** Migrated a Frankfurt-based fintech from on-premises infrastructure to AWS using EC2, RDS, and S3, reducing infrastructure costs by 50% and improving delivery through CI/CD; also 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.

Project

Docstring Verifier — AI-Assisted VS Code Extension https://github.com/akrisanov/docstring-verifier

- Built a proof-of-concept of TypeScript extension that validates Python docstrings using AST-based analysis and 11 deterministic rules, then uses GitHub Copilot through the VS Code Language Model API to generate context-aware Quick Fixes; achieved 86% test coverage.

Education

Bryansk State University Bryansk, Russia
Master's in Applied Mathematics & Information Science *2006 — 2011*

Skills

- **AI Infrastructure & LLM Inference:** vLLM, Kubernetes, NVIDIA H200/H100, distributed model serving, continuous batching, tensor/data parallelism, KV-cache management, performance profiling, benchmarking, latency and throughput optimization
- **Platform Reliability:** AI gateways, model routing, rate limiting, fallbacks, multi-data-center resilience, SLOs, capacity planning, load testing, Prometheus, Grafana, OpenTelemetry, DCGM
- **Backend & Distributed Systems:** Python, FastAPI, Go, C#, .NET, TypeScript, REST APIs, event-driven architecture, PostgreSQL, Redis, ClickHouse
- **Ops & Cloud:** Linux, Docker, Terraform, Ansible, AWS, GCP
- **Technical Leadership:** system architecture, RFCs/ADRs, cross-team technical direction, mentoring