

Anton Alekseev

Senior MLOps / ML Platform Engineer

Remote · CET (UTC+1)

Telegram @antonaleks · alekseev.ap605@gmail.com · LinkedIn · GitHub · Habr · getMentor · ml-infra.pro

GPUs in Kubernetes | Inference | ML platforms | LLM | IaC

SUMMARY

ML infrastructure / MLOps engineer with **9 years in engineering**, 4 of them in ML infrastructure. Builds ML and inference platforms on Kubernetes: a Kubeflow ML platform for **~1,000 users** on hundreds of GPUs and ML and LLM inference on KServe serving **over 10k RPS**. Focus: GPU sharing and scheduling, LLM inference (vLLM, SGLang, llm-d, disaggregated prefill/decode), autoscaling, IaC. **Program expert of the MLOps course** at Yandex Practicum; speaker at HighLoad++, DevOops and PyCon.

9 yrs	~1,000	>10k	15→60%	16	20+
in engineering	ML platform users	inference RPS	GPU utilization	talks	consultations

EXPERIENCE

Avito · Senior MLOps / SRE · 08.2025 - present

15→60% GPU utilization | >10k RPS | days→minutes model to production | 2× fewer pipeline resources

- Raised GPU utilization from ~15% to ~60% on a Kubeflow ML platform (~1,000 monthly active users) running on hundreds of bare-metal Ampere, Hopper and Blackwell GPUs (HGX, NVLink, InfiniBand): Kueue team quotas with checkpointed preemption, topology-aware scheduling, DRA, node health checks and remediation, DCGM dashboards.
- Own the KServe inference platform for ML and LLM end to end, including on-call and SLOs: over 10k RPS across multiple data centers behind a LiteLLM gateway with per-team quotas; cut model time to production from days to minutes.
- Run distributed LLM serving on llm-d with vLLM and SGLang: DP/TP/PP over NVLink and InfiniBand (NCCL), Ray clusters for multi-node inference, disaggregated prefill/decode with a distributed KV cache, KV-cache-aware routing via Gateway API Inference Extension (InferencePool, endpoint picker).
- Tuned vLLM, SGLang and Triton with AIPerf: up to 2.2x throughput and 67% lower inter-token latency through runtime, scheduler and attention-backend tuning; FP8, AWQ and GPTQ quantization, prefix caching, multi-LoRA serving, autoscaling on queue and KV-cache metrics.
- Scaled the platform to ~70k pipeline runs and up to 3k distributed training runs a month across 3,000+ namespaces; wrote Kubeflow plugins and Kubernetes operators in Go; team upstream patches to Kubeflow Pipelines halved pipeline resources.

Yandex Practicum · Program expert, MLOps course (part-time) · 07.2025 - present

10 modules | 5+ authors

- Built a 10-module program, wrote specs and reviewed materials for a team of 5+ authors, designed the course infrastructure and optimized per-student cloud costs.

Selectel · MLOps Engineer · 12.2022 - 08.2025

8 h→24 min platform delivery to a client

- Cut ML platform delivery to a client from 8 hours to 24 minutes: on-demand platform on Kubernetes (ClearML, JupyterHub, Keycloak, Istio) with Terraform and GitLab CI instead of manual steps.
- Built a Triton-based inference platform for cloud customers, replacing Seldon Core: canary, autoscaling, scale to zero, weight caching.
- Ran GPU sharing R&D (MIG, time-slicing, MPS, dynamic MIG) and shipped automatic MIG and time-slicing partitioning to the ML platform; launched a CUDA image marketplace.
- Sized infrastructure for LLMs: VRAM math, GPU choice, load testing, vLLM vs SGLang comparison, quantization.

Gazprom Neft · Software Engineer → R&D Engineer → DevOps Engineer · 02.2019 - 12.2022

- Moved on-prem infrastructure in two regions (OpenStack, Proxmox) to IaC with Terraform, Ansible and GitLab CI; set up JupyterHub for the data science team.
- Developed forecasting models (CatBoost) and IIoT equipment diagnostics services (MQTT, Kafka).

Independent consultant · 02.2025 - present

20+ consultations

- 20+ one-on-one consultations on ML infrastructure via getMentor: ML platform architecture, GPUs in Kubernetes, sizing infrastructure for LLMs.
- Bank (consulting and custom development): deployed an ML platform on the bank's servers with Helm charts (ClearML, Feast, JupyterHub, KServe) and built an automated training and rollout pipeline for a churn model.

Earlier (2017 - 2019): Asulab, MES systems and Siemens/B&R PLCs; a1qa, Python test automation.

SKILLS

GPUs in Kubernetes: NVIDIA GPU Operator, MIG/MPS/time-slicing, Kueue, DRA, KAI Scheduler, Volcano; HGX, NVLink, InfiniBand
Inference: KServe, llm-d, vLLM, SGLang, Triton, Ray, LiteLLM, Gateway API Inference Extension, Knative, Istio, AIPerf
ML platforms: Kubeflow (Pipelines, Trainer), ClearML, MLflow, JupyterHub, Airflow, Feast
Infrastructure: Kubernetes, Helm, Terraform, Ansible, GitLab CI, Argo CD, Flux CD, Argo Workflows, OpenStack, Vault
Observability: Prometheus, VictoriaMetrics, Grafana, DCGM
Languages: Python, Go, Bash

SELECTED TALKS

Talks are in Russian.

- ML platforms for large companies: from DevBoxes to a platform for a thousand users · infra.conf 2026
- Platformizing AI systems: problems of large and small companies · MLechny Put 2026
- GPU inference in K8s: acceleration, sharing and scaling without the pain · DevOops 2025
- NVIDIA Triton Inference Server: building production ML without developers · HighLoad++ 2024
- No pain, no reinvented wheels: getting the most out of GPUs in Kubernetes with resource sharing and easy driver setup · PyCon Russia 2024
- DevOps without ClickOps: how Terraform and GitLab made our infrastructure deploys 20x faster · DevOpsConf 2024

All 16 talks: ml-infra.pro/en/talks/ · 12 articles on Habr · an article shortlisted for Habr's «Technotext 2023»

COMMUNITY

20+ one-on-one consultations on getMentor; Telegram channel on AI and ML infrastructure (t.me/mlops_infra, ~1,000 subscribers).

LANGUAGES

Russian - native · English - professional working proficiency (B2)