

Антон Алексеев

Senior MLOps / ML Platform Engineer

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

GPU в Kubernetes | Инференс | ML-платформы | LLM | IaC

КРАТКО

MLOps-инженер, **9 лет в инженерии**, из них 4 года в ML-инфраструктуре. Строю ML- и inference-платформы на Kubernetes: ML-платформа на Kubeflow для **~1000 пользователей** на сотнях GPU, инференс ML и LLM на KServe с нагрузкой **больше 10 тыс. RPS**. Специализация: шеринг и планирование GPU, LLM-инференс (vLLM, SGLang, llm-d, дезагрегация prefill/decode), автоскейлинг, IaC. **Программный эксперт курса MLOps** в Яндекс Практикуме, спикер HighLoad++, DevOops и PyCon.

9 лет	~1000	>10k	15→60%	16	20+
в инженерии	пользователей ML-платформы	RPS инференса	утилизация GPU	выступлений	консультаций

ОПЫТ

Авито · Senior MLOps / SRE · 08.2025 - н.в.

15→60% утилизация GPU | >10k RPS | дни→минуты вывод модели в прод | 2× меньше ресурсов на пайплайны

- Поднял утилизацию GPU с ~15% до ~60% на ML-платформе Kubeflow (~1000 активных пользователей в месяц, сотни GPU Ampere, Hopper и Blackwell): шеринг GPU через Kueue с квотами команд и вытеснением с чекпоинтом, topology-aware scheduling для HGX по NVLink и InfiniBand, DRA, дашборды утилизации на DCGM.
- Сократил вывод модели в прод с дней до минут: спроектировал и разрабатываю inference-платформу на KServe (Knative и Standard) для ML и LLM в геораспределённом кластере, больше 10 тыс. RPS.
- Запустил распределённый LLM-инференс на стеке llm-d с vLLM и SGLang: DP/TP/PP, дезагрегация prefill/decode с распределённым KV-cache, маршрутизация через Gateway API Inference Extension (InferencePool, EPP). Подбирал параметры Triton, vLLM и SGLang нагрузочными тестами в AIPerf.
- Масштабировали платформу до ~70 тыс. ранов пайплайнов и до 3 тыс. запусков распределённого обучения в месяц при 3000+ namespace.
- Снизили в 2 раза расход ресурсов Kubernetes на пайплайны: доработки Kubeflow Pipelines командой с отправкой в апстрим; интеграция Vault в платформу.

Яндекс Практикум · программный эксперт курса «MLOps для разработки и мониторинга моделей» (part-time) · 07.2025 - н.в.

10 модулей | 5+ авторов

- Собрал программу из 10 модулей, писал ТЗ и ревьюил материалы команды из 5+ авторов, спроектировал инфраструктуру курса и оптимизировал облачные затраты на студента.

Selectel · MLOps Engineer · 12.2022 - 08.2025

8 ч→24 мин развёртывание платформы клиенту

- Сократил развёртывание ML-платформы клиенту с 8 часов до 24 минут: платформа по требованию на Kubernetes (ClearML, JupyterHub, Keycloak, Istio), Terraform и GitLab CI вместо ручных шагов.
- Построил inference-платформу на Triton для клиентов облака вместо Seldon Core: canary, автоскейлинг, scale to zero, кеш весов.
- Провёл R&D шеринга GPU (MIG, TimeSlicing, MPS, динамический MIG) и внедрил автоматическую разметку MIG и TimeSlicing в ML-платформу, запустил маркетплейс образов с CUDA.
- Подбирал инфраструктуру под LLM: расчёт VRAM, выбор GPU, нагрузочные тесты, сравнение vLLM и SGLang, квантизация.

Газпром нефть · Software Engineer → R&D Engineer → DevOps Engineer · 02.2019 - 12.2022

- Перевёл on-prem инфраструктуру двух регионов (OpenStack, Proxmox) на IaC: Terraform, Ansible, GitLab CI; развернул JupyterHub для DS-команды.
- Разрабатывал прогнозные модели (CatBoost) и IIoT-сервисы диагностики оборудования (MQTT, Kafka).

Независимый консультант · 02.2025 - н.в.

20+ консультаций

- 20+ индивидуальных консультаций по ML-инфраструктуре на getMentor: архитектура ML-платформ, GPU в Kubernetes, подбор инфраструктуры под LLM.
- Банк (консалтинг и заказная разработка): развернул ML-платформу на серверах банка Helm-чартами (ClearML, Feast, JupyterHub, KServe) и собрал автоматический пайплайн обучения и выкатки модели оттока.

Ранее (2017 - 2019): Asulab, MES-системы и контроллеры Siemens/B&R; a1qa, автотесты на Python.

НАВЫКИ

GPU в Kubernetes: NVIDIA GPU Operator, MIG/MPS/TimeSlicing, Kueue, DRA, KAI Scheduler, Volcano; HGX, NVLink, InfiniBand
Инференс: KServe, llm-d, vLLM, SGLang, Triton, Gateway API Inference Extension, Knative, Istio, AIPerf
ML-платформы: Kubeflow (Pipelines, Trainer), ClearML, MLflow, JupyterHub, Airflow, Feast
Инфраструктура: Kubernetes, Helm, Terraform, Ansible, GitLab CI, ArgoCD, FluxCD, Argo Workflows, OpenStack, Vault
Наблюдаемость: Prometheus, VictoriaMetrics, Grafana, DCGM
Языки: Python, Go, Bash

ИЗБРАННЫЕ ДОКЛАДЫ

- ML-платформы для больших компаний: от DevBox'ов до платформы на тысячу пользователей · infra.conf 2026
- Платформизация ИИ-систем: проблемы больших и малых компаний · MLечный путь 2026
- GPU-инференс в K8s: ускорение, шеринг и скейлинг без боли · DevOops 2025
- Nvidia Triton Inference Server: строим production ML без разработчиков · HighLoad++ 2024
- Без болей и велосипедов: GPU в Kubernetes с максимальной отдачей, шерингом ресурсов и простой настройкой драйверов · PyCon Russia 2024
- DevOps без ClickOps. Как terraform и gitlab ускорили нам деплой инфраструктуры в 20 раз · DevOpsConf 2024

Все 16 выступлений: ml-infra.pro/talks/ · 12 статей на Хабре · статья в шорт-листе «Технотекст 2023»

СООБЩЕСТВО

20+ индивидуальных консультаций на getMentor, Telegram-канал «Инфраструктура для AI и ML» (~1000 подписчиков).

ЯЗЫКИ

Русский - родной · English - B2, рабочий