турбо

ИИ в прод за две недели

как запускать модели без большой команды, простоя и переплат



Илья Васильченко

Директор по технологическому развитию и стратегии

турбо_облако

облако высоких скоростей

60+

облачных сервисов

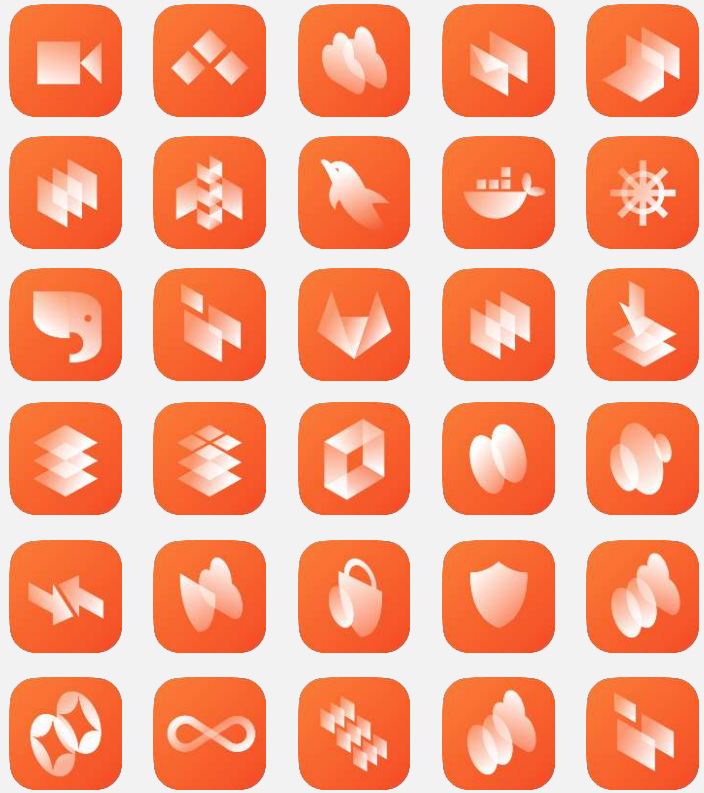
SaaS
Бизнес-приложения

PaaS
Платформенные сервисы

IaaS
Виртуальная инфраструктура и сеть

Информационная безопасность

Профессиональные сервисы



ИИ внедряют, НО эффект измеряют реже

ИССЛЕДОВАНИЕ

95 %

«95% организаций не получают отдачи»

The GenAI Divide

НОВОСТЬ

9 из 10

«Бизнес свернул или отложил 9 из 10 проектов по внедрению генеративного ИИ»

Ведомости

ИССЛЕДОВАНИЕ

>80 %

«Более 80% не видят ощутимого влияния GenAI на EBIT компании»

McKinsey

Как превратить расходы на ИИ в измеримый результат?

ИИ-модель — это больше не редкость

Редкость — умение встроить
её в бизнес-процесс
и довести до продакшена



С какими задачами к нам приходят

29 %

Разработка

Автодополнение в IDE	12 %
Генерация кода	10 %
Ревью и рефакторинг	7 %

19 %

Документы и данные

Анализ документов	7 %
Таблицы, отчётность, перевод	5 %
Таблицы и отчётность	4 %
Перевод документов	3 %

12 %

Прочее

Голос и обработка речи	4 %
Классификация	3 %
Изображения	3 %
HR и онбординг	2 %

19 %

Ассистенты сотрудников

Внутренний чат	8 %
Черновики писем и текстов	6 %
Саммари встреч	5 %

12 %

Клиентская поддержка

Чат-бот первой линии	7 %
Ассистент оператора	5 %

9 %

Агентные системы

Мультишаговые агенты	5 %
Доступ к CRM / ERP	4 %

Задачи остались —
но мы переосмысливаем процессы с нуля

2016–2026	2027
1 Быстрое развертывание тестовых сред и разработка	Использование тестовых сред агентами
2 Организация рабочих мест, включая «удаленку»	Гибридные места с ИИ-ассистентами
3 Автоматизация процессов и внедрение ИИ	Глубокая автоматизация процессов

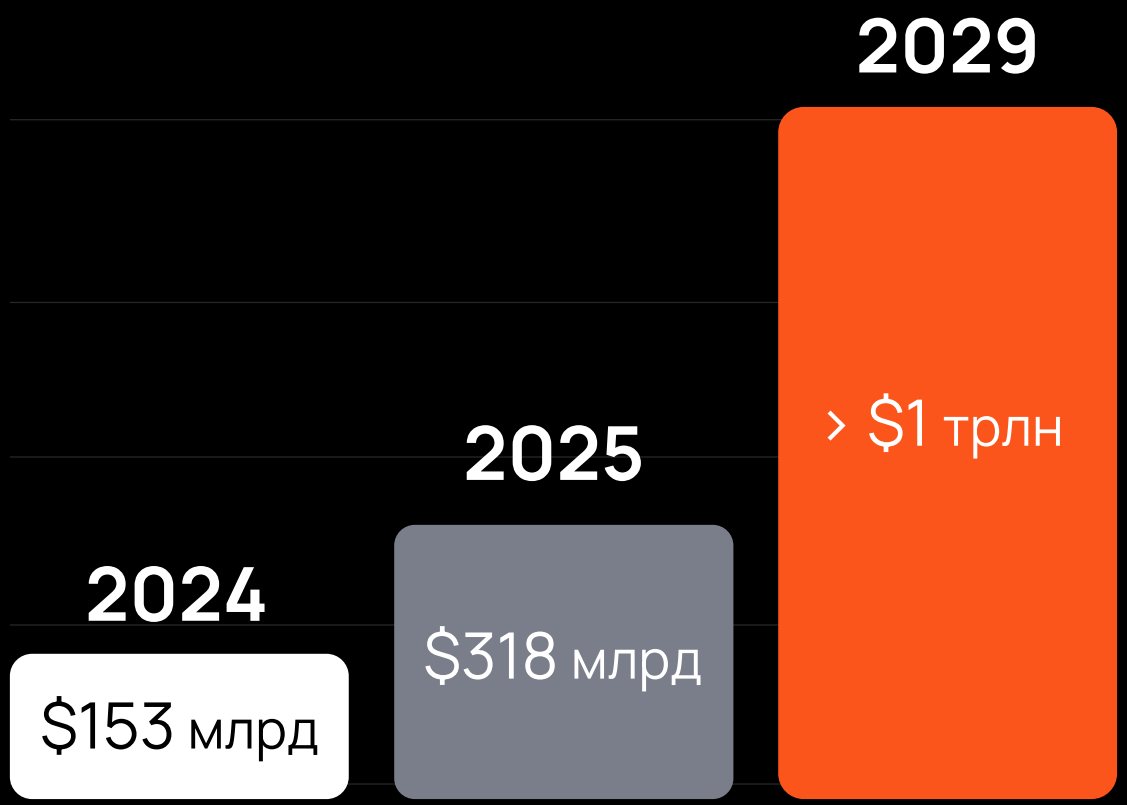


Задачи остались —
но мы переосмысливаем процессы с нуля

2016–2026	2027
4 Резервное копирование и аварийное восстановление	Резервное копирование моделей и весов
5 Платформы для аналитики и скоринга	Сквозная предиктивная аналитика
6 Автоматическая обработка пиковых нагрузок	Балансировка пиковых нагрузок при помощи ИИ



ИИ ломает кривую расходов — переходим от экспериментов к пользе



2026

inference › training

$23,3 млрд **$19 млрд**

AI-optimized IaaS

+96% **$42 млрд**

Источники IDC 2026, Gartner 2026

Покупаем своё железо

Покупаем не только GPU — берём весь стек

8 GPU
рабочий сервер

+8 GPU
аппаратный резерв

+ сеть, диски, кабели, оптика, размещение и эксплуатация

36 мес.
горизонт распределения вложений

720 ч/мес.
владение ресурсом

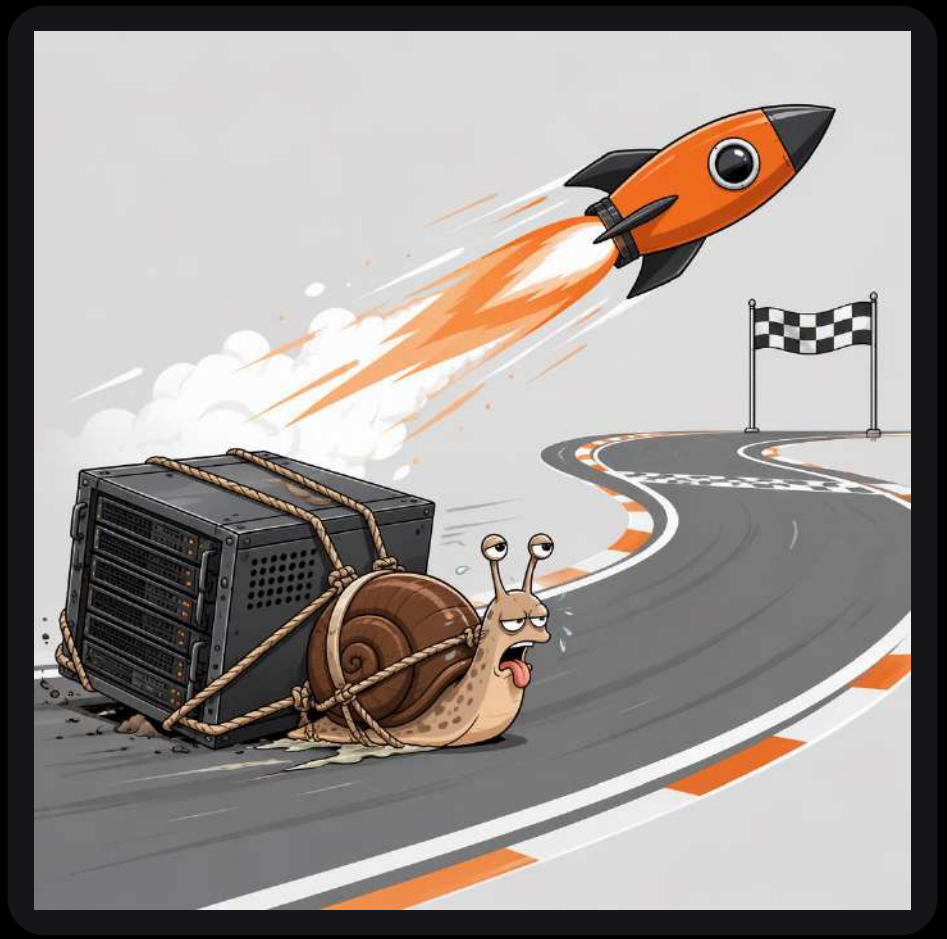
Команда клиента отвечает за:

- Железо и запасные части
- Сеть и связность
- Storage и бэкапы
- Виртуализация / ОС
- Контейнеры и оркестрация
- GPU-драйверы и runtime
- Модель, версии и качество
- Мониторинг, безопасность, восстановление

Зоны ответственности клиента, а не 8 отдельных ставок:
у каждого слоя нужен владелец и процедура восстановления

Зато у нас будет полный контроль над runtime

фраза, которая стоила девять месяцев,
три найма и одного сгоревшего тимлида



Ракета справа уже в проде
Улитка с сервером на спине считает это неспортивным



Сервер с GPU — это ещё не Inference

Аренда железа снимает закупку

НО

...всё остальное

вы строите сами

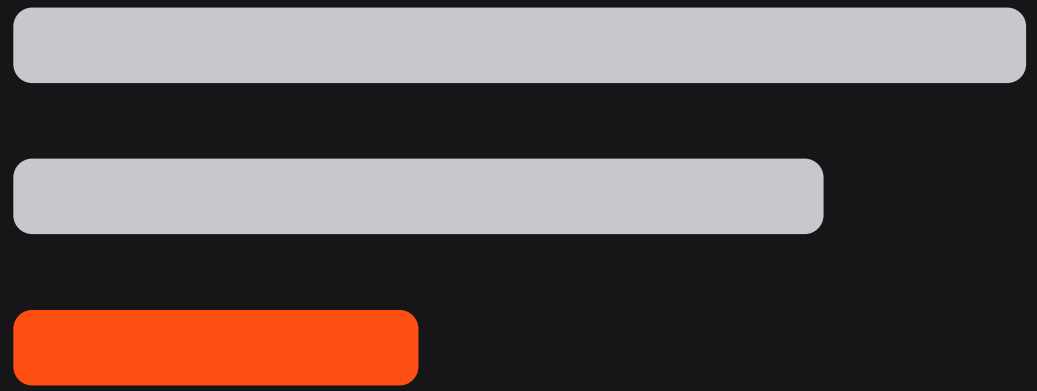
Мы просто арендуем GPU...
Что может пойти не так?

Это последние слова перед тремя
месяцами дежурств и фразы
«у меня локально всё работало»



Смета на месяц работы под нагрузкой

Своё железо	5–10 млн ₽
Физический сервер с GPU	4–8 млн ₽
Inference Platform, 220 ч	2–4 млн ₽



Месячная смета — ещё не цена токена

Что происходит с нагрузкой внутри оплаченного времени?

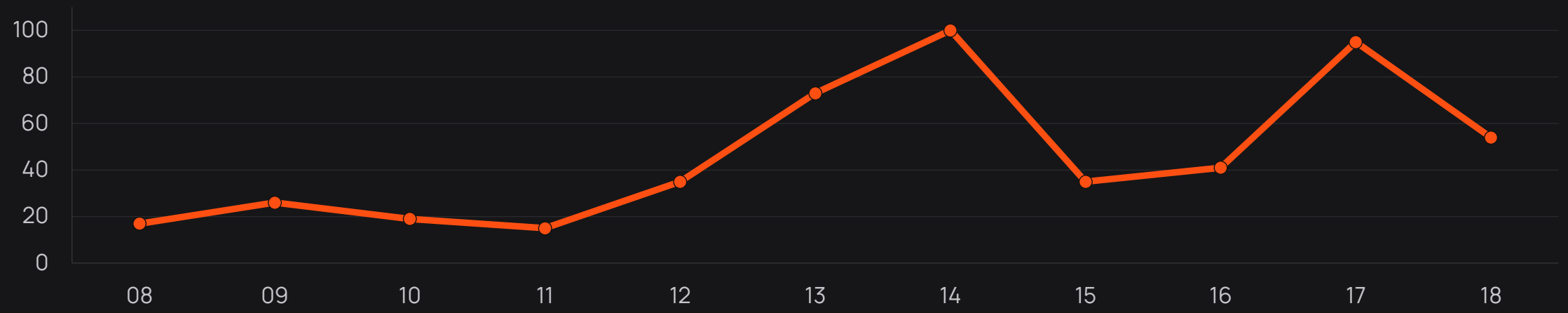
Ресурсы оплатили, сколько токенов выпустили?

46 %
средняя загрузка в окне 08:00–18:00

101 ч
эквивалент полной мощности из 220 ч

×2,2
к фактической цене токена

Нагрузка по часам, рабочее окно 08:00–18:00



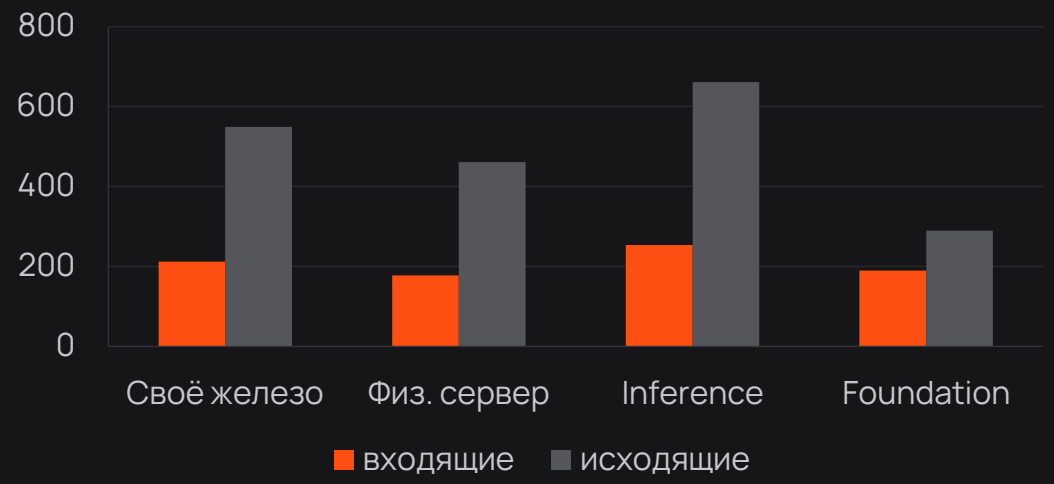
*Загрузка мощности по часам рабочего окна 08:00–18:00. Среднее 46 % — это около 101 часа из 220 оплаченных.

Токен при 46% загрузке стоит иначе

Себестоимость 1 млн токенов при фактической загрузке 46 %, ₽ без НДС

ВАРИАНТ	ВХОДЯЩИЕ	ИСХОДЯЩИЕ
Своё железо	212	549
Физический сервер с GPU	178	461
Inference Platform	254	661
Foundation Model Hub	190	290

Цена 1 млн токенов при загрузке 46 %, ₽



Модель готова. Продакшена всё ещё нет

Модель готова за дни. Инфраструктура — за месяцы.

Прототип

- модель отвечает
- демо работает
- бизнес видит потенциал

3–9

месяцев

между демо и продом

Продакшен

- доступы и ИБ
- GPU и runtime
- масштабирование
- мониторинг и SLA
- резервирование
- эксплуатация 24×7

«Здесь теряется бизнес-время»

демо ИБ · GPU · runtime · масштабирование · мониторинг · SLA · резервирование **прод**

Считать не GPU-часы, а бизнес-результат

Токены — единица потребления. Результат — единица ценности.

Единицы потребления

- токены
- запросы
- GPU-часы

Единицы ценности

- стоимость обработки
- AHT / FRT
- lead time
- число релизов
- высвобождённое время

Полезный результат / потраченный токен

Вы оплатили 220 часов. Модель отработала 101

220 часов	101 час	46 %	×2,2
оплачено	полезной мощности	средняя загрузка	к фактической цене токена

Нагрузка по часам, рабочее окно 08:00–18:00



Оплачено против использования

220 ч · оплачено

101 ч · использовано

×2,2

к фактической цене токена

*Сценарный расчёт для рабочего окна 08:00–18:00. Фактическая экономика зависит от модели, профиля нагрузки и SLA.



Ваш GPU в три часа ночи

Оплачено: 220 часов
Из них 119 часов сервер честно грел воздух
и поддерживал микроклимат в ЦОД

Зато простой получился высокодоступным —
99,95 % времени он стабильно ничего не делал

---

scale-to-zero придумали не от скуки

От IaaS к PaaS: меньше обвязки, быстрее результат

«Эксплуатация модели переходит к платформе» →

IaaS

«Дайте мне GPU»

- BareMetal
- GPU VM / контейнер

PaaS

«Дайте мне endpoint»

- Inference Platform
- Foundation Model Hub

рекомендуемый путь

SaaS

Дайте мне «готовый инструмент»

прикладные ИИ-сервисы

Карты — сырьё. Endpoint — продукт

Железо продаёт ресурс. Платформа продаёт сервис

Своё железо

- CAPEX и резерв
- поставка и размещение
- весь стек и эксплуатация
- постоянная оплата простоя

GPU BareMetal / BM

- быстрый доступ к ресурсу
- runtime и деплой — ваши
- мониторинг модели — ваш
- платите за выделенную мощность

PaaS

Foundation / Inference

- готовый API
- эксплуатация платформы включена
- масштабирование и восстановление
- потребление связано с нагрузкой

Первые два варианта сокращают путь к GPU

Третий — сокращает путь к бизнес-результату

Оставьте себе то, что отличает ваш бизнес

«Здесь создаётся бизнес-ценность»

Под контролем клиента

- данные и доступы
- бизнес-логика
- качество ответа
- интеграции
- лимиты и бюджеты

API

На стороне платформы

- GPU и runtime
- развёртывание
- масштабирование
- обновления
- мониторинг и восстановление

Контроль остаётся
Инфраструктурная рутина уходит

Как выбрать продукт

ЧТО ДЛЯ ВАС КРИТИЧНЕЕ?

ПОЛНЫЙ КОНТРОЛЬ → ГОТОВЫ САМИ ПОДНИМАТЬ ВИРТУАЛИЗАЦИЮ И СТЕК?

- ДА — ПОДНИМАЕМ САМИ
- НЕТ — ГОТОВАЯ VM

СКОРОСТЬ ЗАПУСКА → ПОДХОДИТ МОДЕЛЬ ИЗ ДОСТУПНЫХ?

- НЕТ — СВОЯ МОДЕЛЬ
- ДА — МОДЕЛЬ ИЗ КАТАЛОГА

Продукт	IaaS — Серверы BareMetal	IaaS — Виртуальная инфраструктура	PaaS — Inference Platform	PaaS — Foundation Model Hub
Брать, если	• нагрузка ровная, 24/7 • нужна особая изоляция • виртуализация — ваша	• нужен контроль runtime • готовый VM или контейнер • есть ML/DevOps-команда	• модель своя или редкая • нужен SLA и рост • своей платформы нет	• модель из каталога • нужен быстрый пилот • · оплата за токены
Формат	Выделенный физический сервер с GPU	ВМ или контейнер с GPU	Виртуальный инстанс	LLM-модель
Тариф	За конфигурации	За ресурсы	За инстансы	За токены
Команда	большая ●●●	средняя ●●○	маленькая ●○○	не нужна ○○○

Если хотя бы половина условий неизвестна — начинайте с PaaS: Foundation Model Hub или Inference Platform

Автоматизация рутины для команды внутреннего аудита



8 GPU → **384 ГБ**

Nvidia L40S с 48 ГБ — видеопамяти

Задачи проекта

КЕЙС

1. **Ускорение и глубина аудита**
   Внедрение внутренних ИИ-инструментов
2. **Быстрый изолированный контур**
   Запуск GPU-инфраструктуры без перестройки текущего ИТ-ландшафта

Экономия за счет гибкости

Старт с минимума

Проверили гипотезы на базовых ресурсах без покупки дорогого железа

Точный расчет мощности

Определили реальные потребности модели на практике

Безопасное масштабирование

Взяли мощный стек тогда, когда подтвердилась ценность — нулевой риск потери бюджета

AI INFERENCE

Запуск любых AI-моделей за несколько минут

Передайте модель — платформа развернет окружение и отдаст готовый API endpoint

Hugging Face | S3 | Private Repo | Fine-tuned

Автозагрузка из Hugging Face и S3

Просто укажите источник модели – платформа развернет ее сама

99,95%

SLA

Доступность с перерасчётом стоимости при нарушении

Без своего кластера

Кластер, Kubernetes, GPU, **инженеры — на стороне платформы**

5 Гб

шаг выделения vRAM

GPU-память тарифицируется по факту использования

Scale-to-zero

В простое инстанс не тарифицируется. При трафике поднимается сам

INFERENCE PLATFORM

Подробнее по ссылке

Гб/ч

только за потребление

не нужно резервировать сервер целиком

КЛАСТЕР

128

NVIDIA H200

Гб/мес.

хранения кэша

Модель остается готовой к быстрому запуску

Автомасштабирование

Реплики растут по RPS и параллельным запросам Вы задаёте min и max

Готовый endpoint

Модель, окружение и доступы собираются в готовый API-контур

12 ТБ+

минимальный шаг vRAM

GPU-память тарифицируется по факту

Крупные модели

Multi-node inference и tensor parallelism распределяют модель по серверам.

Совместимость с OpenAI API

Для LLM-моделей достаточно поменять base URL — интеграция без переписывания клиентского кода

от 7,5 ₽

за 1 ГБ памяти в час

Прозрачная тарификация по выделенной GPU-памяти

Inference Platform — для ИИ-стартапа KodaCode

7 млрд
токенов в день



95 тыс.
запросов в сутки

Задачи внедрения

КЕЙС

1. **Отказоустойчивость сервиса**
   Работа ИИ-ассистента для кода без простоев
2. **Масштабирование ×12 (с 10 до 120 тыс.)**
   Подготовка к росту пользователей
3. **Гибкость и комплаенс**
   Уход от фиксированных серверов и локализация данных в РФ

Экономия на кадрах и запуске

Ноль затрат на команду эксплуатации

Не нанимали штат DevOps и SRE-инженеров 24/7 — платформа взяла мониторинг и поддержку инференса на себя

Отказ от простаивающего железа

Ушли от дорогой аренды выделенных серверов с фиксированной оплатой в пользу оплаты по фактической нагрузке

Автомасштабирование под пики

Ресурсы подключаются мгновенно при наплыве запросов и сжимаются при затишье — нет переплаты за простой мощностей

AI Model Hub

Все уже готово и развернуто

Просто выберете модель — платформа развернет окружение и отдаст готовый API endpoint

Обновляемый каталог с актуальными моделями

Не нужно заниматься подбором и настройкой модели

99,95%

SLA

Доступность с перерасчётом стоимости при нарушении

Без своего кластера

Кластер, Kubernetes, GPU, инженеры — на стороне платформы

Qwen

3.8

Наиболее оптимальные модели для офисный сценариев и RAG

Scale-to-zero

В простое инстанс не тарифицируется. При трафике поднимается сам

FOUNDATION MODEL HUB

Подробнее по ссылке

Токены

оплата только за потребление

не нужно резервировать сервер целиком

КЛАСТЕР

H200

NVIDIA

DeepSeek

v4.1 Flash

Дешевый вайбкодинг и агенты

Inference platform

Под капотом со всеми возможностями скейлинга в зависимости от нагрузки

Готовый endpoint

Модель, окружение и доступы собираются в готовый API-контур

GLM 5.3

аналог Claude Opus

Решение Frontier класса

Крупные модели

Multi-node inference и tensor parallelism распределяют модель по серверам

Совместимость с OpenAI API

Для LLM-моделей достаточно поменять base URL — интеграция без переписывания клиентского кода

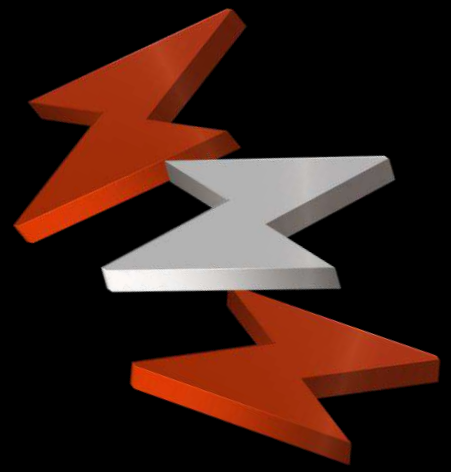
Лимиты по токенам

день/неделя/месяц

Возможность устанавливать индивидуальные лимиты на отдел/департамент/компанию

Умный ассистент техподдержки — на Foundation Model Hub

Более 4 FTE высвобождено



КЕЙС

Задачи внедрения

1. **Ускорение техподдержки за счет внедрения диалогового ИИ**
2. **База знаний в модели**
   Работа с документацией, логами и инцидентами
3. **Снятие рутины с инженеров и сокращение времени поиска решений**

Экономия на кадрах и запуске

Без отдельной команды эксплуатации

Инфраструктура платформы сняла потребность в штате DevOps и инженеров сопровождения

Запуск силами одного ИИ-специалиста

Один эксперт закрыл весь цикл внедрения — от подготовки данных до боевого запуска

Ответ кадровому голоду

Готовый хаб моделей позволил обойтись без найма дефицитных ML-разработчиков

Две недели вместо девяти месяцев

Неделя 1

- Сценарий и KPI
- Данные и доступы
- Выбор Foundation или Inference

Неделя 2

- Интеграция API
- Нагрузочная и качественная проверка
- Лимиты, мониторинг и запуск

ПУТЬ С ТУРБО · 10 РАБОЧИХ ДНЕЙ

Д1	Д2	Д3	Д4	Д5	Д6	Д7	Д8	Д9	Д10

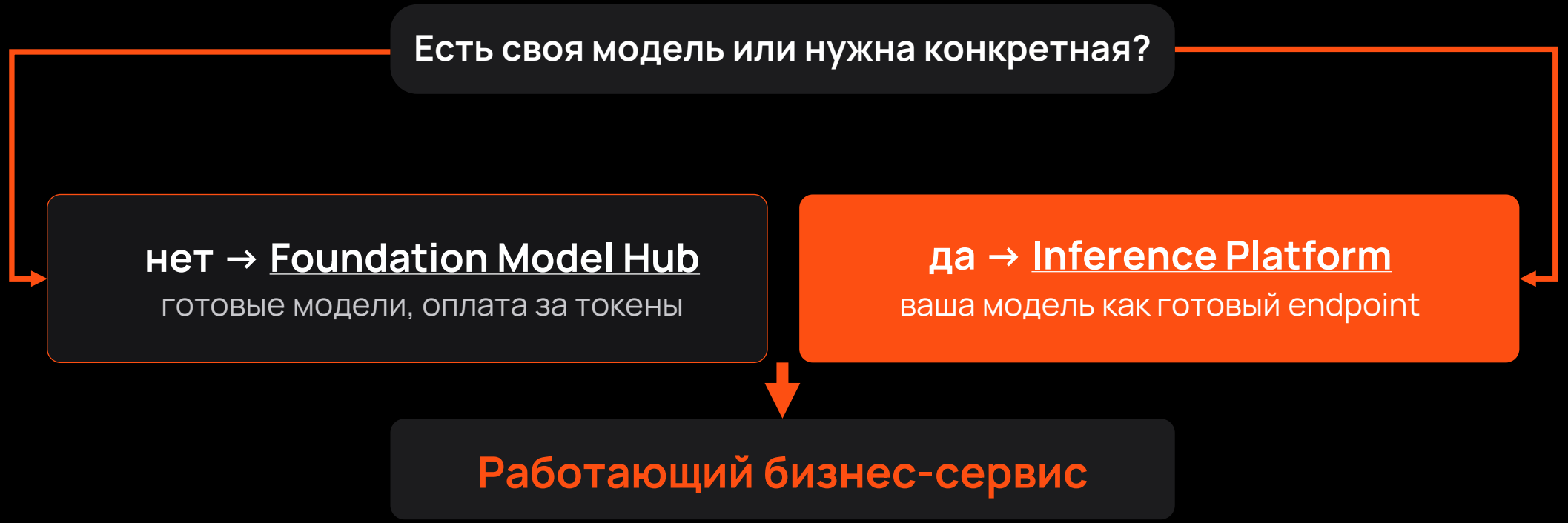
KPI · контур данных · endpoint · интеграция · тест · production

ТРАДИЦИОННЫЙ ПУТЬ · 3–9 МЕСЯЦЕВ

закупка · сборка платформы · найм · согласования

Не «внедрить ИИ за 14 дней», а вывести ограниченный измеримый сценарий в прод

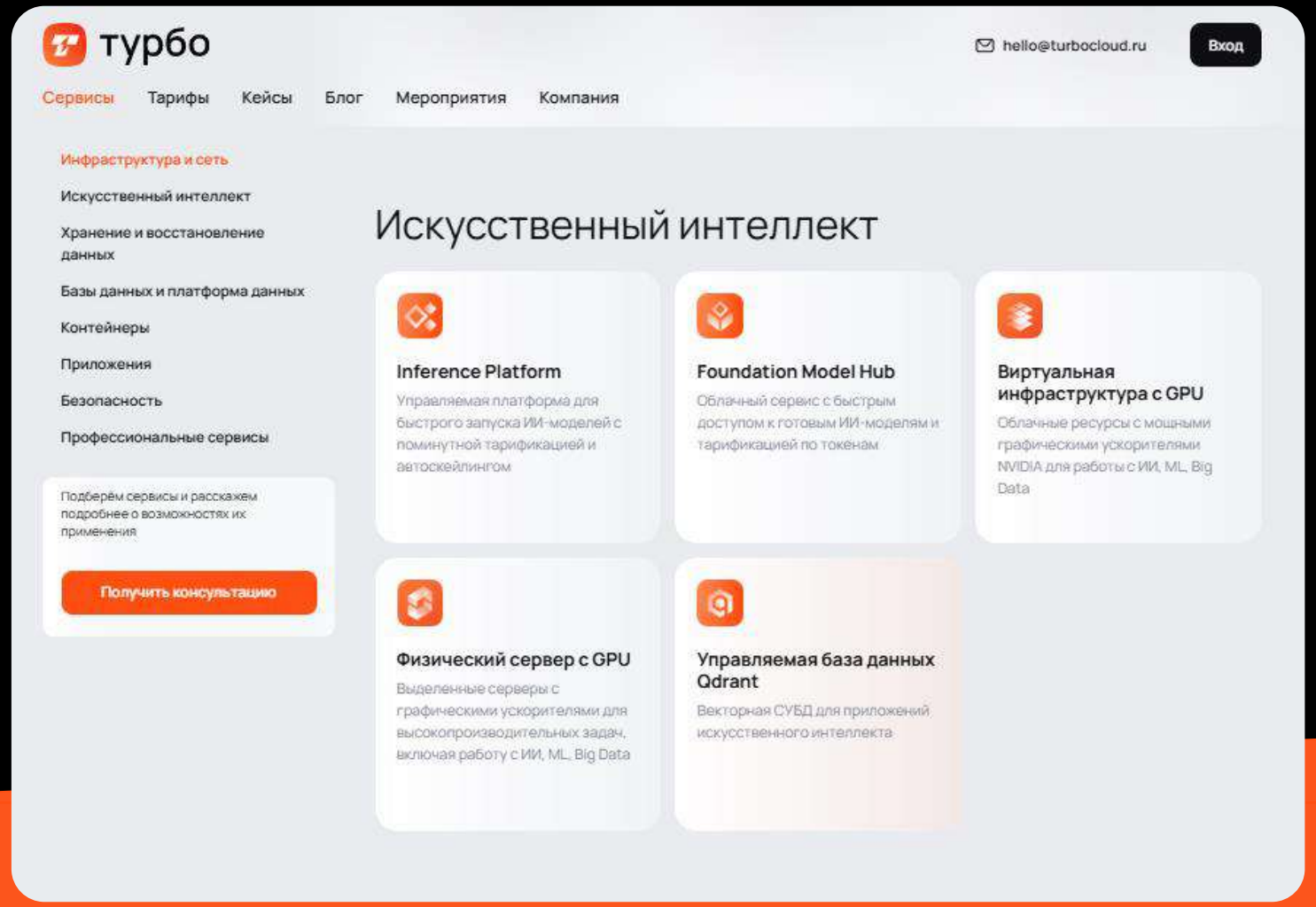
Не стройте ML-платформу,
если ваш бизнес — не ML-платформа



Один сценарий · один KPI · две недели до production

Начните с пилота в «Турбо Облаке»

БЫСТРЫЙ СТАРТ С ТУРБО



TurboCloud.ru