турбо

Технологии роста

как бизнес интегрирует облачные сервисы и ИИ в свои процессы

Илья Васильченко

Директор Центра технологического развития и стратегии

13 лет эволюции

2012
Ростелеком

2013
DataLine
SafeData

2023
РТК-ЦОД
Публичное облако

С 2025

турбо
турбо_облако

турбо_облако

облако высоких скоростей

50+

облачных сервисов

- **SaaS**
  Бизнес-приложения
- **PaaS**
  Платформенные сервисы
- **IaaS**
  Виртуальная инфраструктура и сеть
- **Информационная безопасность**
- **Профессиональные сервисы**

«Облако рядом с клиентом»

20+
площадок

13+
лет в облачном провайдинге

500 000+
vCPU

>40 ПБ
storage

Облачные сервисы на базе геораспределенной сети дата-центров уровня Tier III с отказоустойчивой инфраструктурой

Повседневные задачи бизнеса
2016–2026

Быстрое развертывание тестовых сред и разработка

Резервное копирование и аварийное восстановление

Платформы для аналитики и скоринга

Организация рабочих мест включая «удаленку»

Автоматизация процессов и внедрение ИИ

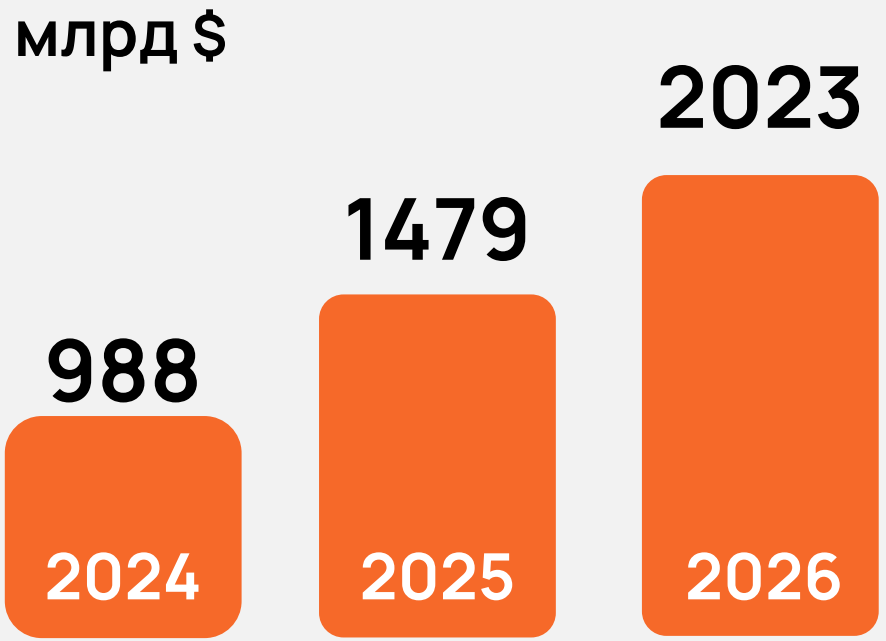
Автоматическая обработка пиковых нагрузок

Тренд

Искусственный интеллект

Инфраструктурные вызовы для бизнеса

Мировой рынок — расходы на ИИ



- AIaaS — ИИ как услуга
- Развитие инфраструктуры
- Интеграция в потребительскую электронику

×10

0,3 Вт
энергоемкость стандартного поискового запроса Google

VS

3,0 Вт
энергоемкость поискового запроса ChatGPT

Как бизнес использует ИИ

- Персонализированое клиентское обслуживание
- Запуск сложных AI-продуктов и исследований
- Обработка и извлечение данных из документов
- Ускорение разработки и аналитики
- Сквозная автоматизация бизнес-процессов
- Прогнозирование и предиктивная аналитика

Проблемы внедрения инфраструктуры с ИИ

Очень дорогое «железо» — необходимы значительные вложения в инфраструктуру (On-premise)

Для оптимизации языковых моделей необходимы дополнительные капитальные вложения

Преимущества облачных сервисов

Быстрый запуск

Экономическая эффективность

Масштабируемость

Сервисы с GPU

	2025 IaaS	2025 IaaS	2026 PaaS	2026 PaaS	2026 SaaS
Продукт	Серверы BareMetal	Виртуальная инфраструктура	Inference Platform	Foundation models	Нейрошлюз
Формат	Выделенный физический сервер с GPU	ВМ или контейнер с GPU	Виртуальный инстанс	LLM-модель	Чат с ИИ-моделями для бизнес-пользователей
Тариф	Конфигурации	За ресурсы	За инстансы	За токены	За пользователя

турбо_облако

решения для ИИ

Физические серверы

BareMetal с GPU

Виртуальная инфраструктура

Виртуальные машины
с виртуализированными GPU

Linux-контейнеры
с доступом к GPU
через технологию MIG

Что выбрать

	BareMetal	Виртуальная машина	Linux контейнеры
Производительность GPU	■■	■	■■■
Гибкое управление конфигурацией	■	■■■	■■
Минимальные накладные расходы	■■■	■	■■
Изоляция от «шума»	■■■	■	■■■
Минимальная стоимость	■■	■■■	■

КЕЙС

Автоматизация рутины — для команды внутреннего аудита

Задачи проекта

- Повысить скорость обработки информации, расширить масштаб анализа и глубину проработки данных. Развитие внутренних ИИ инструментов для аудита
- Для запуска: оперативно развернуть производительную изолированную инфраструктуру с GPU, не перестраивая текущую ИТ-среду

8 GPU Nvidia L40S с 48 ГБ → 384 ГБ видеопамяти

Результат

- Развернута виртуальная инфраструктура, с инструментами для управления, обучения нейросетей и ускоренного вывода результатов: Container Toolkit, CUDA Toolkit, cuDNN и TensorRT
- Настроено высокоскоростное хранилище на базе NVMe/SSD-дисков для быстрой загрузки и обработки данных при обучении моделей
- Решение стало технологическим фундаментом для масштабирования использования ИИ в компании

КЕЙС

ИИ в техподдержке — для команды ИТ-провайдера

Задачи проекта

- Внедрить нейросети в бизнес-процессы техподдержки для роста оперативности и качества сервиса
- Развернуть решение с GPU для внедрения открытых языковых моделей — цифровых помощников, которые понимают естественный язык, анализируют большие объемы техдокументации и генерируют код

в 6 раз
ускорили выполнение

на 70%
сократили время реакции инженеров на инциденты

Результат

- Развернута виртуальная инфраструктура с H200, в изолированном защищенном контуре. Производительность разработчиков выросла в 3-4 раза: на 20-50% времени меньше на рутину
- Структурированная информация об инциденте позволяет немедленно приступить к его устранению
- Цифровым помощникам делегированы: ведение и контроль процесса, участие в анализе логов, подготовка отчетности и рекомендаций

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

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

КЕЙС

Inference Platform — для ИИ-стартапа KodaCode

Задачи проекта

- Обеспечить стабильную работу ИИ-помощника для программистов без простоев
- Уйти от негибкой аренды физических серверов с постоянной оплатой
- Подготовиться к 12-кратному росту аудитории (с 10 до 120 тысяч пользователей)
- Гарантировать локализацию данных в российском контуре для B2B-клиентов

7 млрд токенов обрабатывается в день

95 тыс. запросов пользователей в сутки

Результат

- Модели размещены на Inference Platform с автоматическим управлением GPU-ресурсами — оплата только за фактическое использование
- Самая тяжелая модель запущена в мультинодовом режиме для высокой скорости ответов даже при сложных запросах
- Система автоматически масштабируется: в пиковые часы подключаются дополнительные мощности, при спаде — отключаются

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

3.6 и 3.5

Наиболее оптимальные модели для офисный сценариев и RAG

Scale-to-zero

В простое инстанс не тарифицируется. При трафике поднимается сам

FOUNDATION MODEL HUB

Токены

оплата только за потребление

не нужно резервировать сервер целиком

КЛАСТЕР

H200

NVIDIA

DeepSeek

v4 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 5.2

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

КЕЙС

ИИ-ассистент поддержки — на базе Foundation Model Hub

Задачи внедрения

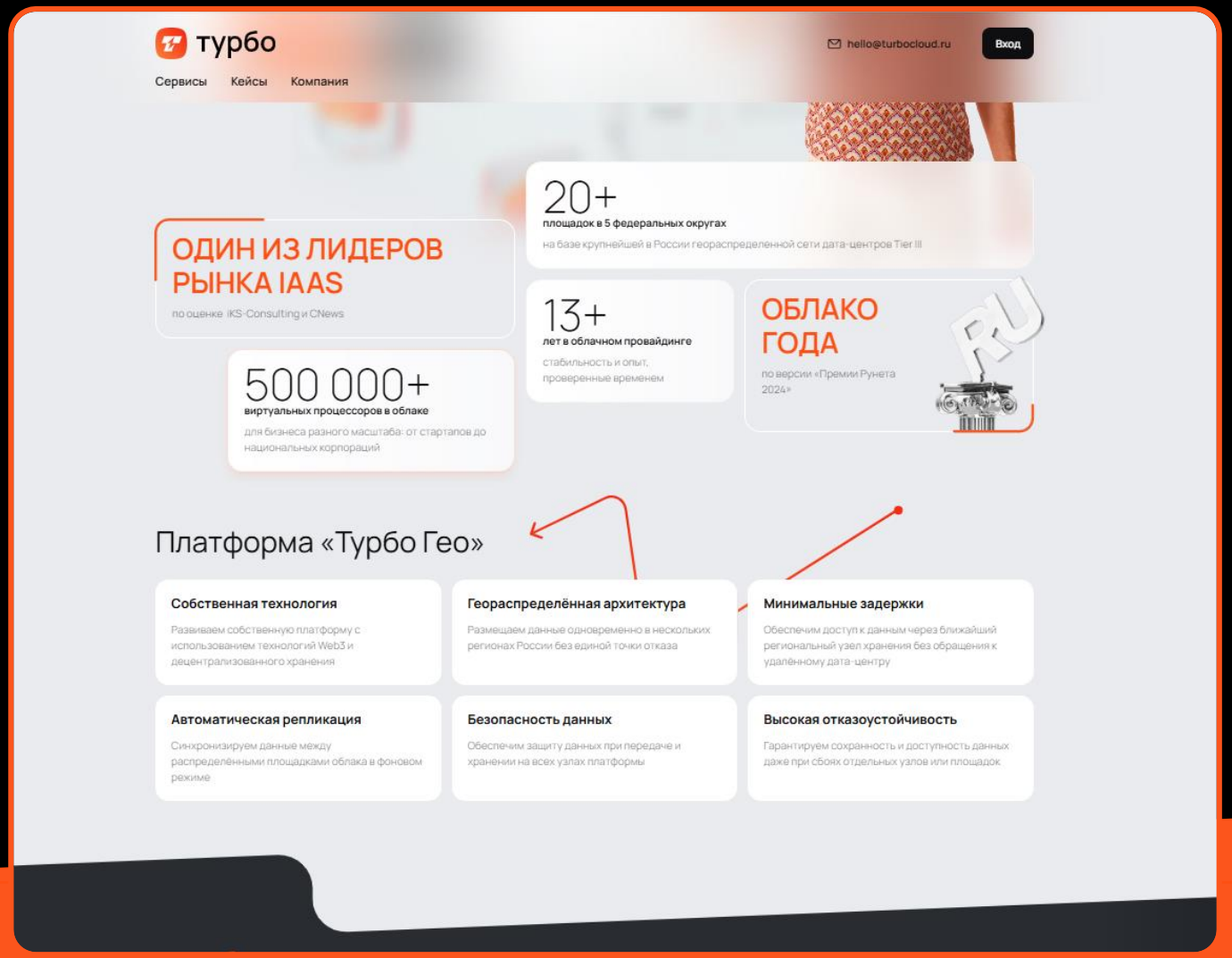
- Интегрировать ИИ в процессы техподдержки для повышения скорости и качества обслуживания клиентов
- Обучить модель работать с технической документацией, логами и историей инцидентов
- Сократить время поиска информации и рутинные операции инженеров

По первой линии поддержки
Высвобождение более 4 FTE

Результат

- Снижение операционной нагрузки на инженеров первой линии
- Ускорение обработки каждой заявки без потери качества
- Клиенты получают развернутые, понятные и персонализированные ответы
- Снижение человеческого фактора при постановке задач и интерпретации логов
- Повышение точности и скорости исполнения инцидентов

БЫСТРЫЙ СТАРТ С ТУРБО



TurboCloud.ru