турбо

Технологии роста

как бизнес интегрирует облачные сервисы и ИИ в свои процессы

Михаил Соколов

Директор по технологиям

Михаил Соколов

Директор по стратегии и технологиям «Турбо Облако»

> Инновации без эмпатии — это просто игрушки для инженеров. Наши технологические решения рождаются из реальных задач и целей клиентов.
>
> Только так технологии обретают силу!

13 лет эволюции

2012
Ростелеком

2013
DataLine
SafeData

2023
РТК-ЦОД
Публичное облако

С 2025

турбо
турбо_облако

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

Как бизнес их решает с помощью облака: от стоек → к IaaS, PaaS, SaaS и AIaaS

2016–2017

- Стойки в ЦОД (Colocation)
- Начало консолидации и виртуализации
- Облако под офисную инфраструктуру

2018–2022

- Стандартизация использования облака регуляторами
- Внедрение PaaS для ускорения разработки
- Повышение отказоустойчивости ИС с помощью IaaS

2023–2026

- GPU как сервис (GPUaaS)
- AI как сервис (AIaaS)
- Гибридное облако — как новая норма
- Разработка, включая микросервисы

турбо_облако

облако высоких скоростей

50+

облачных сервисов

SaaS
Бизнес-приложения

PaaS
Платформенные сервисы

IaaS
Виртуальная инфраструктура и сеть

Информационная безопасность

Профессиональные сервисы

Тренд

Облако рядом

Вызовы географической распределенности бизнеса

Цифровое неравенство близости к облакам в России

«Облако рядом с клиентом»

20+
площадок

13+
лет в облачном провайдинге

500 000+
vCPU

>40 ПБ
storage

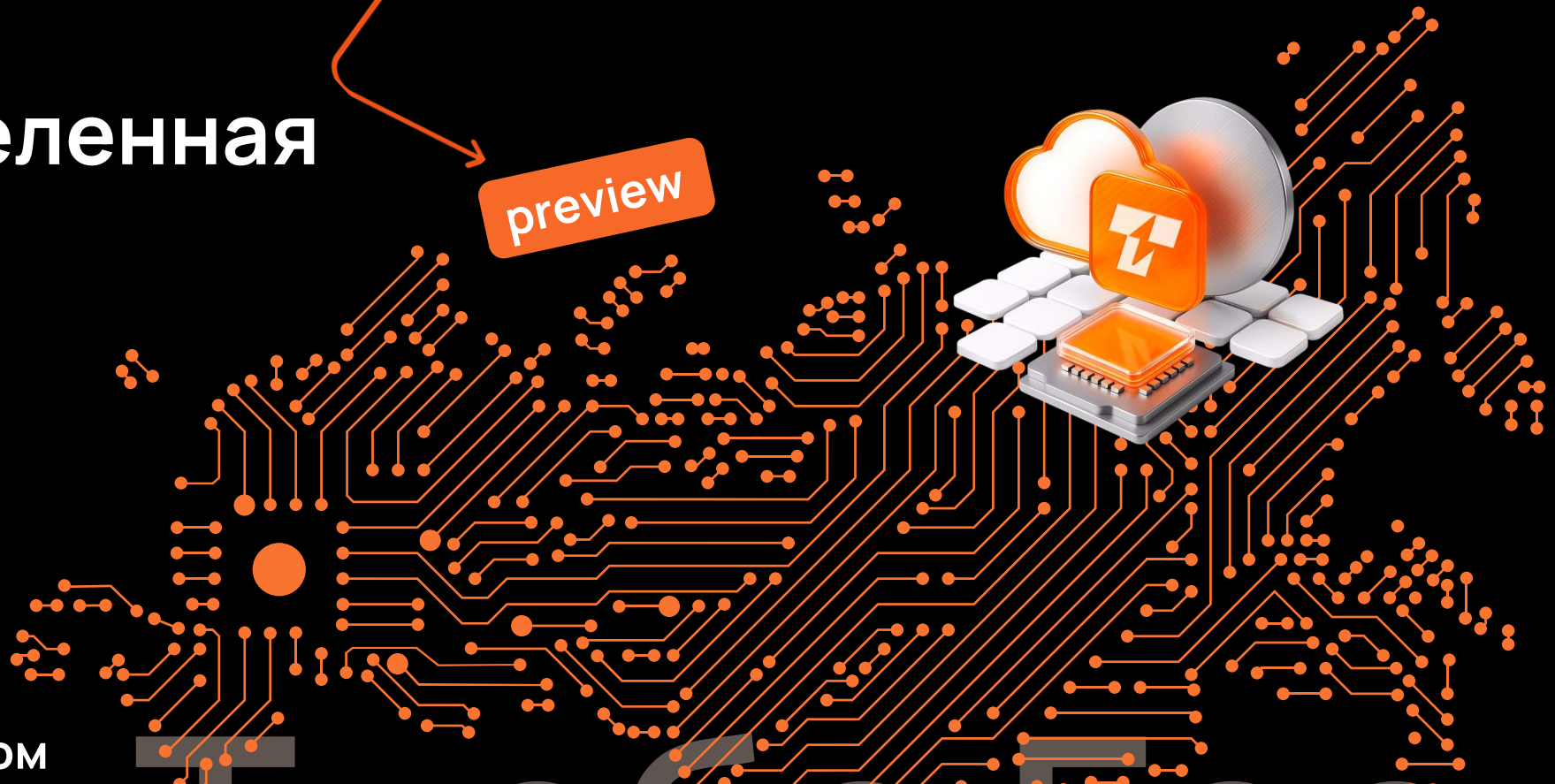
Облачные сервисы на базе геораспределенной сети дата-центров уровня Tier III с отказоустойчивой инфраструктурой

турбо_облако

Быстрый доступ к сервисам
и данным во всех регионах России

Георасределенная платформа

preview



Этап 1→
На базе сети
ЦОДов РТК-ЦОД

Этап 2→
Масштабирование
на узлы сети Ростелеком

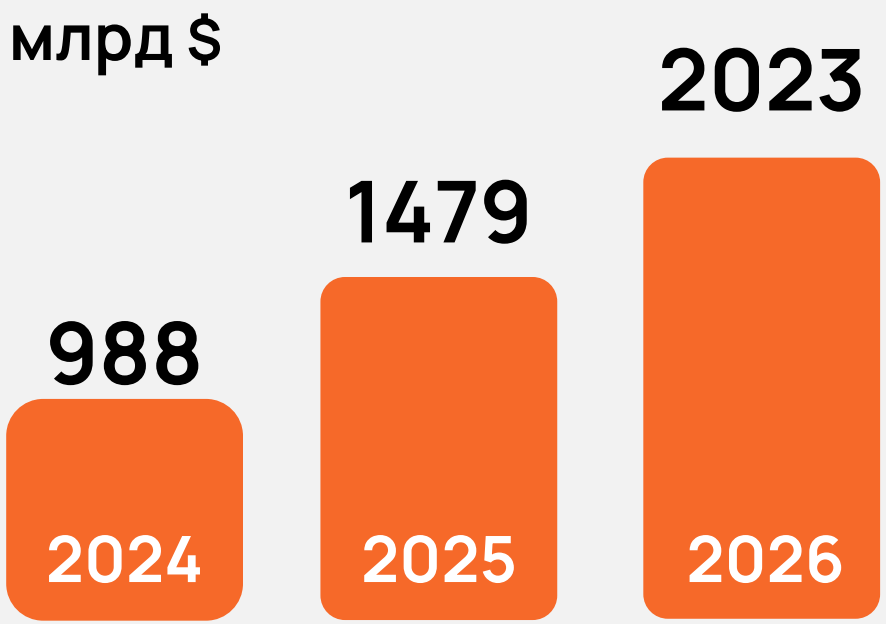
Турбо Гео

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	АПРЕЛЬ 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	ML-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

Готовы обсудить партнерство или первыми узнать о запуске пилотов?

Соколов Михаил

Turbo Cloud.ru