
Conews
FORUM

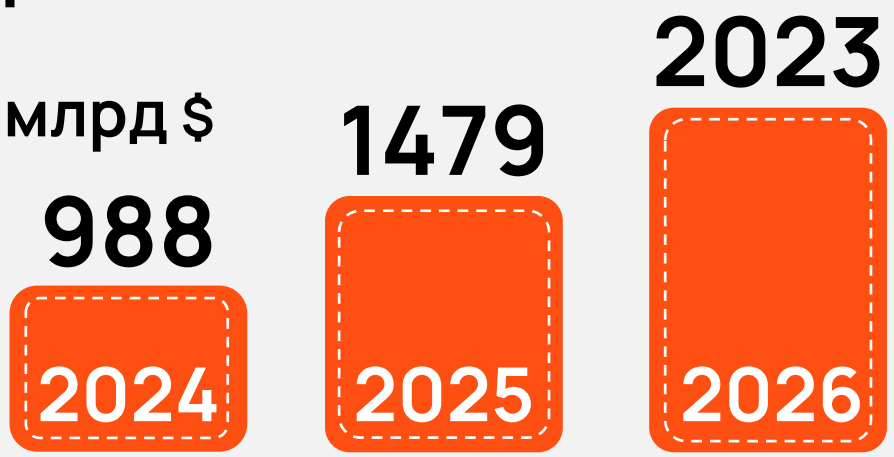
турбо_облако

Сложность vs Доступность:

Как облако демократизирует искусственный интеллект

Михаил Соколов

Мировой рынок
расходы на ИИ



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

интересно

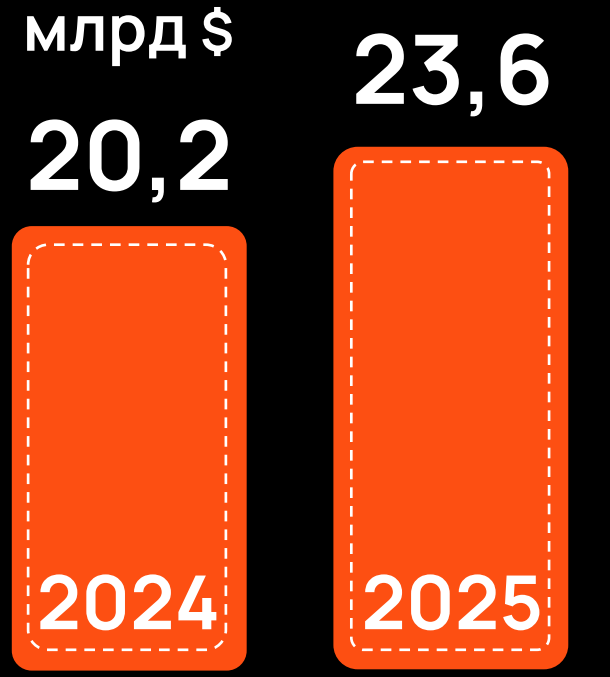
×10

0,3 Вт
энергоемкость стандартного поискового запроса Google

VS

3,0 Вт
энергоемкость поискового запроса ChatGPT

Стремительный рост ИИ → драйвер рынка видеокарт



18,2%
CAGR до 2032

Операторы дата-центров наращивают мощности

- Обучение больших языковых моделей (LLM)
- Поддержка ресурсоемких ИИ-приложений
- Инференс

Сценарии использования

- 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

Тренд

ИИ рядом

Вызовы географической распределенности

турбо_облако

качество доступа

Географически распределенная платформа

Q1 2026

Создание инфраструктурных сервисов для продуктов AI/ML

Предпосылки

→ Потребность рынка в вычислениях на GPU

82
клиентских запроса в ‘22–’23 на 1000+ GPU разных классов

→ У большинства облачных игроков продукты с GPU уже запущены

Инициатива

2024

→ **Предоставление вычислительных мощностей на базе GPU из облака с возможностью выбора оптимальной конфигурации: ML, 3D, VDI**

→ **Создание MVP сервисов AI/ML**

Анализ опыта запуска GPU

Выводы

Клиентам нужны
не «сырые» GPU,
а виртуализированные
и преднастроенные
для работы среды

2025 →

Сервисы

- Созданную инфраструктуру с GPU масштабируем
- Запускаем новый сервис «Виртуальные машины с GPU»
- Разворачиваем высокоуровневые сервисы для ML/AI

Сервисы для AI

Продукт	Серверы BareMetal	Виртуальная инфраструктура	Inference	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	За конфигурацию сервера	За виртуальные ресурсы	За инстансы	За токены	За пользователя
Доступно	от 150 340 ₽ / мес	от 11 900 ₽ / мес	Скоро	Скоро	Скоро

Что выбрать

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

Выделенные серверы BareMetal

Без виртуализации для ресурсоемких задач, включая серверы с графическими картами для нейросетей, работы с графикой и машинного обучения

- Современные конфигурации серверов
- Широкий выбор операционных систем
- Инструменты для удаленного управления сервером
- Единый портал управления облаком и серверами BareMetal
- Аппаратное резервирование компонентов сервера
- Возможна доукомплектация базовой конфигурации
- Поддержка NVLink
- Нет переподписки по PCIe
- Доступна поддержка функционала виртуальных дисков

- Сетевая связанность с облачной инфраструктурой
- В составе сервера могут быть предоставлены GPU-карты

Выделенные серверы BareMetal

	Сервер Medium	Сервер Medium Plus	Сервер Medium Pro	Сервер Medium Ultra	Сервер Large GPU
CPU	AMD EPYC 7313 32 ядра, 3.0 GHz	Intel Xeon Gold 6240 36 ядер, 2.6 GHz	AMD EPYC 7443 48 ядер, 2.85 GHz	Intel Xeon Gold 6248R 48 ядер, 3.0 GHz	AMD EPYC 7543 64 ядра, 2.8 GHz
RAM	от 256 ГБ	от 512 ГБ	от 512 ГБ	от 1024 ГБ	от 512 ГБ
Хранение	SAS / SATA / NVMe SSD	SATA / SAS / SSD	SAS / SATA / NVMe SSD	SATA / SAS / SSD	SAS / SATA / NVMe SSD
Кол-во GPU	До 1-й на сервер	До 2-х на сервер	До 2-х на сервер	До 2-х на сервер	До 8-ми на сервер
Поддерживаемые GPU	L4, A6000	L4, A6000, A100	L4, L40S, A6000, A100, H100	L4, A6000, A100	L4, L40S, A6000, A100, H100
+Комплектация RAM, HDD, SSD	Да	Да	Да	Да	Да

Виртуальная инфраструктура с GPU

Позволяет делиться ресурсами физического GPU между несколькими ВМ или контейнерами, обеспечивая гибкость и эффективность без владения дорогостоящим оборудованием

- Выделенный профиль GPU на каждую ВМ/контейнер
- Отдельный образ для Data Science с популярными инструментами и фреймворками
- Масштабирование в пару кликов
- Инструменты для удаленного управления ВМ
- Подготовленные образы с оптимизацией под GPU
- Возможность заказа ВМ и контейнера с несколькими GPU
- Современное железо с DDR5 и PCIe Gen5
- Единый портал управления

- TensorFlow, PyTorch и другие инструменты для вашей работы
- Гарантированная производительность без влияния «шумных соседей»

Конфигурации виртуальных машин с GPU

	Профиль 1	Профиль 2	Профиль 3	Профиль 4	Профиль 5	Профиль 6	Профиль 7
GPU	L4	L4	L4	L4	L40S	L40S	L40S
Профиль GPU	3 ГБ	6 ГБ	12 ГБ	24 ГБ	12 ГБ	24 ГБ	48 ГБ
vCPU, ядра	4	8	16	32	8	16	32
RAM	16 ГБ	32 ГБ	64 ГБ	128 ГБ	48 ГБ	96 ГБ	192 ГБ
SSD	60 ГБ	60 ГБ	60 ГБ	60 ГБ	60 ГБ	60 ГБ	60 ГБ

Конфигурации Linux-контейнеров с GPU

	Профиль 1	Профиль 2	Профиль 3	Профиль 4	Профиль 5
GPU	H200	H200	H200	H200	H200
Профиль MIG	1g.35gb	2g.35gb	3g.71gb	4g.71gb	7g.141gb
vCPU	14	14	28	28	56
RAM	64 ГБ	64 ГБ	128 ГБ	128 ГБ	256 ГБ
SSD	60 ГБ	60 ГБ	60 ГБ	60 ГБ	60 ГБ

Новые сервисы

2026 +

турбо_облако

облако высоких скоростей

50+

облачных сервисов

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

турбо_облако

Public preview — Q1 2026

скоро

Нейро-шлюз

Private Preview — уже сейчас

Онлайн-платформа, где можно работать с множеством средств и сервисов AI в едином рабочем пространстве

турбо_облако

Public preview — Q1 2026

скоро

Inference platform

Позволяет создавать инстансы с необходимыми библиотеками и инструментами для гибкого управления ML-моделями

турбо_облако

Public preview — Q2 2026

скоро

Foundation models

Сервис обеспечивает доступ к актуальным ML-моделям, расположенным на доверенной инфраструктуре провайдера, с тарификацией по токенам

БЫСТРЫЙ СТАРТ С ТУРБО

Соколов Михаил

Turbo Cloud.ru