

 турбо_облако

GPU как сервис –
повышаем эффективность
бизнеса без лишних затрат





Елена Синегубкина

**Отвечает за развитие
IaaS-сервисов**

 турбо_облако

→ **Облако**
высоких скоростей

 турбо_облако

50+

облачных
сервисов

облако высоких скоростей



SaaS

Бизнес-приложения



PaaS

Платформенные
сервисы



IaaS

Виртуальная
инфраструктура и сеть



Информационная
безопасность



Профессиональные
сервисы



Стремительный рост ИИ → драйвер рынка видеокарт



млрд \$

20,2

2024

23,6

2025

18,2%
CAGR до 2032

Операторы дата-центров
наращивают мощности

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

Сценарии использования



 Персонализированное 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

Преимущества облачных сервисов



Быстрый
запуск

Экономическая
эффективность

Масштабируемость



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



Производительность
GPU

Гибкое управление
конфигурацией

Минимальные
накладные расходы

Изоляция от «шума»

Минимальная
стоимость

BareMetal



Виртуальная
машина



Linux
контейнеры



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 VM или контейнерами, обеспечивая гибкость и эффективность без владения дорогостоящим оборудованием

- Выделенный профиль GPU на каждую VM/контейнер
- Отдельный образ для Data Science с популярными инструментами и фреймворками
- Масштабирование в пару кликов
- Инструменты для удаленного управления VM
- Подготовленные образы с оптимизацией под GPU
- Возможность заказа VM 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

Особенности предоставления сервиса



Конфигурации

- Виртуальные машины сконфигурированы с учетом наиболее оптимальной производительности
- GPU — с учетом рекомендаций Nvidia

Варианты тарификации

- Allocated — за весь пул ресурсов в месяц
- PAYGO — посуточная или почасовую





Кейсы



→ графические ускорители в бизнесе

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 —————→ 384 ГБ
Nvidia L40S с 48 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

ИИ в техподдержке — для команды ИТ-провайдера



Задачи проекта

- Внедрить нейросетевые технологии в бизнес-процессы техподдержки продуктов для роста оперативности и качества сервиса
- Развернуть решение с использованием GPU для внедрения открытых языковых моделей — цифровых помощников, которые понимают естественный язык, анализируют большие объемы технической документации и генерируют программный код

в 6 раз
ускорили
выполнение

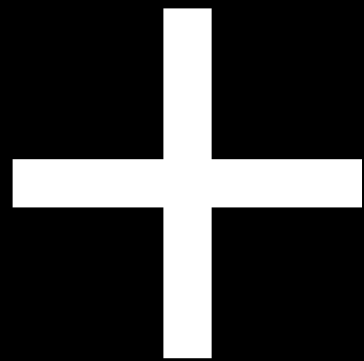
на 70%
сократили время реакции
инженеров на инциденты

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

Новые сервисы

2026



 НейрошлюзQ1
2026

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

 Inference platformQ1
2026

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

 Foundation modelsQ2
2026

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

Как протестировать сервисы с GPU бесплатно



1 ⚡

Обратиться к вашему
аккаунт-менеджеру
или подать заявку
с сайта TurboCloud.ru

2 ⚡

Выбрать
конфигурацию —
виртуальную
машину или
контейнер

3 ⚡

Получить доступ
на **7 дней**
и протестировать
сервис с GPU
без ограничения
функционала

БЫСТРЫЙ
СТАРТ С ТУРБО



TurboCloud.ru

 **турбо**

hello@turbocloud.ru **Вход**

Сервисы Компания

Виртуальная инфраструктура с GPU

Облачные ресурсы с мощными графическими картами NVIDIA для ускорения решения задач ML, Big Data и сложной графики

[Оставить заявку](#)



Сервис для высокопроизводительных вычислений

Искусственный интеллект и машинное обучение

Мощные GPU ускоряют разработку и развертывание моделей, что позволяет быстрее запускать ИИ-продукты на рынок и экономить ресурсы



Научные исследования

Высокопроизводительные ресурсы для параллельных вычислений, числовых расчетов, анализа данных и моделирования физических процессов



Визуализация и рендеринг

Стабильная работа со сложными графическими задачами, 3D-моделями и видео без затрат на дорогостоящие локальные станции



