FORSITE
СУПЕРКОМПЬЮТЕРЫ

СПЕЦИФИКАЦИЯ

Сервер Devbox AI

Сервер DEVBOX AI — это гибкое и масштабируемое решение, адаптируемое под задачи генеративного контента, анализа данных и обучения моделей ИИ.

В зависимости от ваших задач вы можете выбрать конфигурацию с GPU, которая подходит как для инференса (например, RTX 5080 или RTX 4090), так и для обучения моделей (например, RTX 5090 или RTX 6000 Ada).

Процессор AMD Ryzen Threadripper PRO 7995WX: 96 ядер и 192 потока с базовой частотой 2,5 ГГц и возможностью разгона до 5,1 ГГц, обеспечивает высокую производительность, необходимую для интенсивных вычислительных задач в области AI и LLM.

Поддержка большого объема оперативной памяти позволяет эффективно работать с крупными наборами данных и сложными моделями, гарантируя плавную и бесперебойную работу даже при максимальных нагрузках.

Выбирая DEVBOX AI, вы получаете надежный инструмент, способный удовлетворить потребности современных AI-разработок, обеспечивая баланс между производительностью, универсальностью и стоимостью.

Особенности Devbox AI

- Выбор GPU под ваши задачи;
- Оптимизировано для AI, ML, LLM и HPC-задач;
- Качественное охлаждение основных компонентов системы;
- Гибкость конфигурации и масштабируемость;
- Совместимость с TensorFlow, PyTorch, JAX, CUDA и другими AI-фреймворками;
- Работает круглосуточно 24/7 без перегрева и падения производительности.

Наши преимущества

15-ти летний опыт работы с GPU решениями

Многоэтапное тестирование компонентов

Высокая производительность

Собственные R&D офисы

Сервис и поддержка в России

Собственное производство и сборочная линия в РФ

до	до	до	до
6 GPU **564 Гб** памяти	**96** ядер **192** потока	**2 ТБ** RAM (ОЗУ)	**4** NVME SSD

Идеальное решение для:

- Разработчиков, работающих с LLM и NLP, которым необходимы мощные и доступные вычислительные ресурсы;
- Стартапов в сфере искусственного интеллекта, ищущих эффективные и экономичные решения для ускорения разработки и вывода продуктов на рынок.

Параметры	Devbox AI
Шасси	19” шасси 6,5U
Процессор	AMD Ryzen™ Threadripper™ PRO 7000 Series
Число ядер / потоков	До 96 ядер / 192 потока
Тактовая частота	Базовая – 2,5 ГГц, буст – до 5,1 ГГц
Чипсет	AMD WRX90
Разъем CPU (Socket)	sWRX9
Совместимые GPU	**до 4 x GPU (на выбор) :** · Nvidia RTX 5080 16 GB · Nvidia RTX 4090 24 GB · Nvidia RTX 5090 32 GB · NVIDIA H200 NVL 141 GB с поддержкой NVLink **до 6 x GPU (на выбор) :** · AMD Pro W7800 AI TOP 32GB · Nvidia RTX 5000 Ada 32GB · Nvidia RTX 6000 Ada 48GB · AMD Pro W7900 AI TOP 48GB
Максимальный объем RAM (ОЗУ)	до 2 ТБ (8 слотов DIMM)
Слоты расширения	7× PCIe 5.0 x16
Система хранения	4× M.2 NVMe PCIe 5.0 (2280/22110) + 2× 2.5» SATA SSD
Сетевые интерфейсы	2× 10Gb Ethernet (Intel X710)
RAID-контроллер	Поддержка RAID 0, 1, 5, 10 (NVMe и SATA)
Охлаждение	9 вентиляторов 80x38 мм с ШИМ-управлением и частотой вращения 12000 об/мин
Разъемы I/O на задней панели	2×10GbE, 1xBMC, 2×USB 4 40Gbs (Type-C), 6×USB 3.2 Gen2 (Type-A), 1× VGA
Блок питания	80 PLUS® Titanium CRPS 3200 Вт
Габариты шасси	650 × 435 × 290 мм (глубина × ширина × высота)
Операционная система	Ubuntu, Astra Linux, Windows
Гарантия	12 месяцев

Сценарии применения Devbox AI:

- **Обучение и дообучение больших языковых моделей (LLM)**
  Обучение и дообучение языковых моделей (GPT, LLaMA, Falcon) под специфические задачи, например, юридический анализ, медицина или техническая поддержка.

- **Интерактивные чат-боты и виртуальные ассистенты**
  Создание голосовых и текстовых ботов, работающих в режиме реального времени с минимальной задержкой.

- **Глубокий анализ данных и прогнозирование (Big Data & AI Analytics)**
  Автоматизация бизнес-процессов, предсказательное моделирование, выявление аномалий в больших наборах данных.

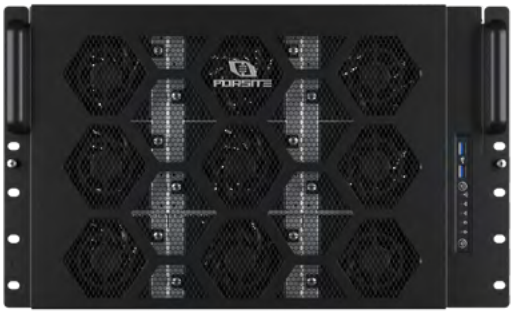
- **Разработка и тестирование моделей компьютерного зрения (CV)**
  Распознавание лиц, анализ видео, детекция аномалий на производствах, автономные транспортные системы.

- **Генерация и рендеринг 3D-графики и анимации**
  Создание фотореалистичных 3D-моделей и анимаций в реальном времени, упрощение работы художников.

- **Генерация музыкального и звукового контента с AI**
  Использование ИИ для автоматического создания музыки и звуковых эффектов (пример: Riffusion, Google MusicLM).

- **Гиперреалистичная генерация персонажей и V-Tubing**
  Создание аватаров и анимированных персонажей на основе видеозахвата.

- **Медицинская диагностика и биоинформатика**
  Автоматический анализ рентгеновских снимков, CT/MRI-изображений, прогнозирование лекарственных взаимодействий.



Devbox AI как инструмент для работы с LLM моделями

Семейство LLaMA (Large Language Model Application) модели, находят применение в широком спектре задач, от простых утилитарных до сложных специализированных приложений.

Рассмотрим примеры использования моделей различного размера, соответствующие задачи и типы клиентов.

1. Малые модели (например, LLaMA 7B):

Задача: Автоматическая генерация текстов.

Пример использования: Создание статей, документов и других текстовых материалов на основе заданных тем.

Тип клиента: Малые и средние предприятия, нуждающиеся в автоматизации создания контента для маркетинговых материалов, блогов или внутренних документов.

2. Средние модели (например, LLaMA 13B):

Задача: Анализ и суммирование больших объемов текста.

Пример использования: Интеллектуальный анализ больших текстовых данных, извлечение важной информации и ее обработка.

Тип клиента: Исследовательские организации или аналитические компании, обрабатывающие большие объемы текстовых данных для получения инсайтов и принятия решений.

3. Крупные модели (например, LLaMA 70B и более):

Задача: Машинный перевод и сложные языковые задачи.

Пример использования: Машинный перевод текста с одного языка на другой.

Тип клиента: Крупные технологические компании или международные корпорации, требующие высококачественного машинного перевода и обработки многоязычных данных.

LLaMA 7B:

- **Без квантизации** (FP16): Требует около 13 ГБ VRAM. Может быть запущена на любом из указанных GPU без распределения нагрузки.
- **С квантизацией** до 4 бит: Размер модели уменьшается до примерно 3,9 ГБ, что позволяет запускать ее даже на GPU с меньшим объемом памяти.

LLaMA 13B:

- **Без квантизации** (FP16): Требует около 26 ГБ VRAM. Может быть запущена на GPU с 32 ГБ памяти (RTX 5090) или на двух GPU с 16 ГБ (RTX 5080) при использовании распределения нагрузки.
- **С квантизацией** до 4 бит: Размер модели уменьшается, что позволяет запускать ее на GPU с меньшим объемом памяти.

LLaMA 70B:

- **Без квантизации (FP16):** Требует около 160 ГБ VRAM. Для запуска потребуется распределение модели на все четыре GPU с 48 ГБ памяти (RTX 6000 Ada), что в сумме даст 192 ГБ VRAM.
- **С квантизацией до 4 бит:** Требования к VRAM снижаются, что может позволить запуск на нескольких GPU с меньшим объемом памяти.

Рекомендации:

- **Квантизация:** Применение методов квантизации (например, до 4 бит) позволяет значительно снизить требования к VRAM, что расширяет возможности вашего сервера для запуска более крупных моделей.
- **Распределение нагрузки:** Использование нескольких GPU для распределения модели позволяет эффективно использовать доступные ресурсы и запускать более крупные модели, чем это возможно на одном GPU.

Таким образом, сервер Devbox AI способен эффективно работать с моделями LLaMA до 70B параметров при условии использования методов оптимизации, таких как квантизация и распределение нагрузки между GPU.