Сервер для AI-инференса (2× NVIDIA A800)

📌 Назначение

Готовое решение для запуска моделей типа ChatGPT, LLaMA, StableLM, генерации текста, голосовых ассистентов, внутренних ИИ-систем в компании. Идеально подходит для компаний, которым важно:

- не зависеть от облаков и внешних API
- держать данные внутри организации
- получить ИИ-инфраструктуру под ключ с установкой и поддержкой

⚙️ Конфигурация оборудования

- **Форм-фактор:** 2U (стойка)
- **CPU:** 2× Intel Xeon Silver
- **ОЗУ:** 256 GB ECC DDR4
- **Накопители:** 2× NVMe 2TB
- **GPU:** 2× NVIDIA A800 40GB PCIe
- **Сеть:** 2× 10GbE (Intel X550-T2 или Mellanox ConnectX-4)
- **Удалённое управление:** BMC с KVM/IPMI

💻 Предустановленное ПО

- Ubuntu Server 22.04 LTS
- NVIDIA CUDA Toolkit
- Docker
- TorchServe (API-платформа для моделей PyTorch)
- HuggingFace Transformers
- Мониторинг: Prometheus + Grafana (по запросу)

💼 Сервис и сопровождение

- Первичная настройка и тестирование включены
- Поддержка по e-mail/Telegram — 3 месяца бесплатно
- Опционально: расширенная гарантия на 3 или 5 лет
- Возможна аренда сервера для теста на 14 дней

💰 Стоимость

- **Цена (с DDP Москва + предустановкой):** 869 676 ₽
- Гарантия 3 года: +50 784 ₽
- Гарантия 5 лет: +88 872 ₽
- Настройка и сопровождение (если отдельно): 76 176 ₽

🧠 Почему выбирают нас?

- Работаем напрямую с ODM (Inspur, Supermicro, Gigabyte)
- Собираем серверы под задачи заказчика
- Поддерживаем доставку и сопровождение по всей РФ
- Предлагаем альтернативу облакам с понятной экономикой

📦 Условия поставки

- Срок отгрузки: 2–3 недели с момента оплаты
- Поставка DDP (включено всё: таможня, логистика, сертификаты)
- Доступен договор поставки с юр. лицом, оплата с НДС