Годовикова 9с17
7 495 234 2222
info@skala-

Описание функциональных характеристик программного обеспечения для электронно-вычислительных машин «Скала^р Спектр ИИ»

Москва
2026

Годовикова 9с17
7 495 234 2222
info@skala-

Содержание:

Годовикова 9с17
7 495 234 2222
info@skala-

Описание функциональных характеристик

«Скала^р Спектр ИИ» - это платформа, устанавливаемая на ПАК МИИ и предназначенная для централизованного управления ресурсами ПАК (CPU/RAM/GPU) в сценариях использования ИИ-моделей, приложений с ИИ-функциональностью, агентских систем, а также рабочих столов.

«Скала^р Спектр ИИ» позволяет повысить эффективность использования GPU-карт в рамках ПАК, соотнеся вид нагрузки и часть ресурса карты. Также «Скала^р Спектр ИИ» сокращает время подготовки ПАК для начала работы с моделями, приложениями, агентами и рабочими столами, а также устраняет необходимость в установке и конфигурации дополнительного специализированного ПО поверх ОС.

Ключевые задачи, решаемые «Скала^р Спектр ИИ»:

- Распределение ресурса, включая GPU-карты, между контейнерами с ИИ-нагрузками, единым образом для карт от разных вендоров
- Мониторинг нагрузки на GPU-карты с возможностью прогнозирования потребностей, исходя из профиля загрузки
- Запуск на ПАК ИИ-моделей как из локального репозитория, так и открытых ресурсов с минимальными настройками и созданием эндпоинтов
- Учет используемых моделями токенов для целей биллинга
- Запуск преднастроенных рабочих мест на базе контейнеров для сотрудников, которым требуется использование GPU
- Запуск на ПАК MCP/A2A-серверов и других приложений на базе контейнеров и публикация эндпоинтов
- Разграничение доступа пользователей к исполняемым ИИ-нагрузкам
- Каталогизация ИИ-сервисов, развернутых на ПАК (модели, агенты, контейнеры)
- Упрощение эксплуатации и настройки ПАК посредством использования агентского режима

«Скала^р Спектр ИИ» предназначен для DevOps/MLOps/AIOps инженеров и для ML/Data Science- специалистов.

Годовикова 9с17
7 495 234 2222
info@skala-

Основные сценарии использования

Основными сценариями использования «Спектр ИИ» являются:

1. Развертывание и предоставление доступа к ИИ-моделям и MCP/A2A-серверам

- Пользователь сценария: DevOps/MLOps-специалист или ML/DataScience-инженер (единицы-десятки сотрудников)
- Основные потребители сценария: инженеры, которым необходим эндопинт развернутого сервиса не посредственно для использования или интеграции с корпоративными приложениями (десятки-сотни сотрудников)
- Ключевые требования:
    - Наличие каталога моделей/MCP/A2A-серверов с возможностью запуска в виде контейнера. При этом модели могут быть размещены как в корпоративном репозитории, так и быть представленными ссылками на HuggingFace.
    - Возможность конфигурации GPU-ресурса, выделенного для запускаемого контейнера
    - Наличие чата для проверки работы запущенной модели
    - Возможность подключения средств фильтрации промптов

2. Развертывание рабочих столов и приложений на базе готовых образов

- Пользователь сценария: DevOps/MLOps-специалист или ML/DataScience-инженер (единицы-десятки сотрудников)
- Основной потребитель сценария: любые сотрудники организации, которым требуется рабочий стол или доступ к развернутому приложению (в пределах 150 человек – определяется параметрами конкретного ПАК и требования образов)
- Ключевые требования:
    - Возможность использования образов из репозитория Docker
    - Средства конфигурации ресурсов, включая GPU, выделяемых для запускаемого контейнера
    - Установка лимитов используемого ресурса

- Сохранение пользовательского профиля и его повторное использование при создании новых рабочих столов
- Интеграция с корпоративным SSO

3. Конфигурация параметров хостов / кластеров

- Пользователь сценария: DevOps/MLOps-специалист (единицы-десятки сотрудников)
- Основной потребитель сценария: DevOps/MLOps-специалист (единицы-десятки сотрудников)
- Ключевые требования
  - Подключение рабочих узлов
  - Подключение / конфигурация Kubernetes-кластеров

4. Мониторинг использования ресурсов ПАК с возможностью прогнозирования

- Пользователь сценария: DevOps/MLOps-специалист (единицы-десятки сотрудников)
- Основной потребитель сценария: DevOps/MLOps-специалист (единицы-десятки сотрудников)
- Ключевые требования
  - Отображение загрузки RAM/CPU/GPU
  - Статистика по токенам (input/output) и обращениям к моделям, позволяющая организовать учет для биллинга
  - Возможность прогнозирования GPU-ресурса, исходя из истории генерации токенов
  - Интеграция с внешними системами мониторинга

5. Запуск безопасных / диагностических операций через ИИ-ассистента ПАК

- Пользователь сценария: DevOps/MLOps-специалист (единицы-десятки сотрудников)
- Основной потребитель сценария: DevOps/MLOps-специалист (единицы-десятки сотрудников)

Годовикова 9с17
7 495 234 2222
info@skala-

- Ключевые требования
    - Возможность выполнить сервисную операцию в режиме чата с запуском агентского сценария
    - Возможность сбора и отображения данных логов и метрик в режиме чата с запуском агентского сценария

Метрики

В «Скала^р Спектр ИИ» определяются следующие метрики (список будет расширяться):

- Количество созданных / запущенных контейнеров
- Общее количество пользователей
- Количество пользователей, одновременно использующих рабочие столы / приложения (контейнеры)
- Распределение GPU между созданными нагрузками
- Общее количество обращений к API моделей
- Общее количество обращений к API A2A-MCP-серверов
- Общее количество completion / prompt токенов
- Количество обращений к каждой запущенной модели
- Общее количество обращений к каждому API/A2A-MCP-серверов
- Количество completion / prompt токенов каждой запущенной модели
- Средняя загрузка CPU / RAM / GPU
- Максимальная загрузка CPU / RAM / GPU
- Количество используемых GPU
- Общее количество тензорных ядер
- Количество созданных рабочих узлов