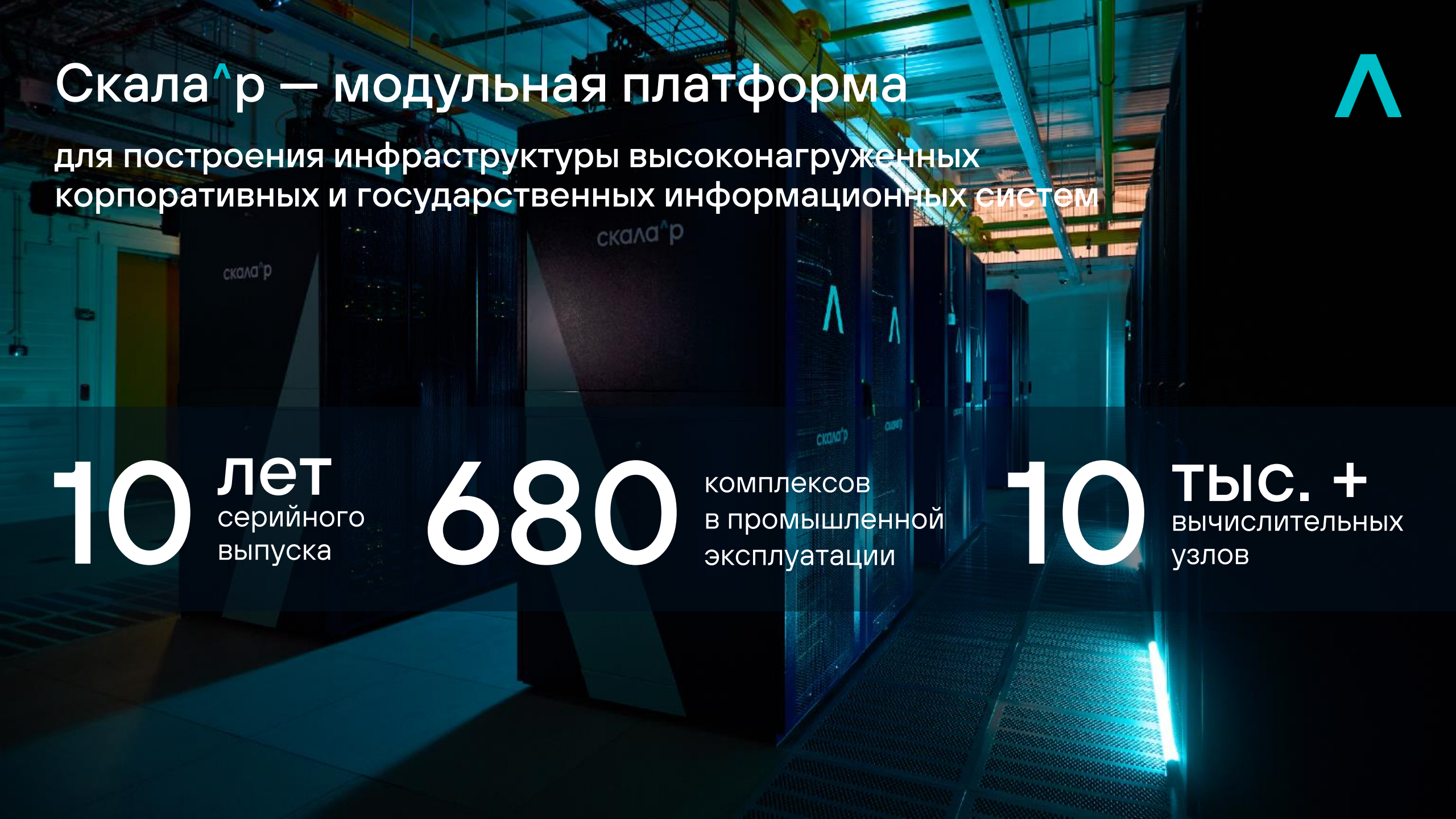
Машина Искусственного Интеллекта Скала^р



Скала^р — модульная платформа

для построения инфраструктуры высоконагруженных корпоративных и государственных информационных систем

10 лет серийного выпуска

680 комплексов в промышленной эксплуатации

10 тыс. + вычислительных узлов

Продуктовые направления Скала^р

решения для высоконагруженных корпоративных и государственных систем

Динамическая инфраструктура

Машины динамической инфраструктуры Скала^р МДИ

на основе решений BASIS для создания динамической конвергентной и гиперконвергентной инфраструктуры ЦОД и виртуальных рабочих мест пользователей

Высокопроизводительные базы данных

Машины баз данных Скала^р МБД

на основе решений Postgres Pro для замены Oracle Exadata в высоконагруженных системах с обеспечением высокой доступности и сохранности критически важных данных

Инфраструктура для ИИ

Машина искусственного интеллекта Скала^р

на основе оптимизированного программно-аппаратного стека для максимальной производительности при работе с моделями ИИ

Управление большими данными

Машины больших данных Скала^р МБД.8

на основе решений ARENADATA и PICODATA для создания инфраструктуры хранения, преобразования, аналитической, статистической обработки данных , а также распределенных вычислений

Интеллектуальное хранение данных

Машины хранения данных Скала^р МХД

на основе технологии объектного хранения S3 для геораспределенных катастрофоустойчивых систем с сотнями миллионов объектов различного типа и обеспечения быстрого доступа к ним

- Использование опыта технологических лидеров (гиперскейлеров)
- Использование самых зрелых и перспективных технологий в кооперации с технологическими лидерами российского рынка в каждом из сегментов

Модульная платформа Скала^р

Использование опыта технологических лидеров — гиперскейлеров

Единый принцип модульной компоновки и платформенный подход

Единая облачная система управления сервисами

IaaS

PaaS

DBaaS

Единая система управления ресурсами и эксплуатацией

Разделение ресурсов

Мультитенантность

Автоматизация

Модульная платформа

Динамическая инфраструктура

- Серверная виртуализация **Dynamix Standard**
- Гиперконвергенция **vCore**
- Контейнеризация приложений **Deckhouse**
- Частное облако **Dynamix Enterprise**
- Виртуализация рабочих мест **WorkPlace**

Динамическая инфраструктура

Инфраструктура управления данными

- Монолитная СУБД **Postgres Pro**
- Шардированная СУБД **Postgres Pro**
- Резидентская обработка **Tarantool**
- Гибридная СУБД **YandexDB**

Транзакционная обработка

- Массивно-параллельная обработка **Greenplum**
- Аналитическая обработка **ClickHouse**
- Неструктурированные данные **Hadoop**
- Потоковая обработка **Kafka и NiFi**

Большие данные

- Объектное хранилище **S3**

- Резервное копирование
- Искусственный интеллект

Специализир. **решения**

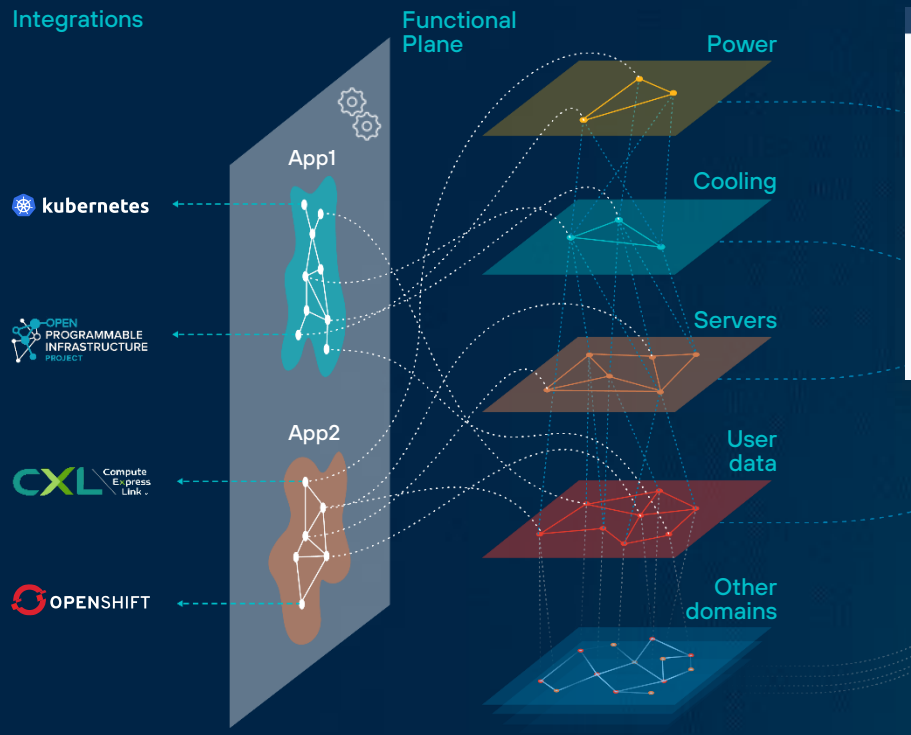
- Автоматизированные банковские системы
- АСУТП

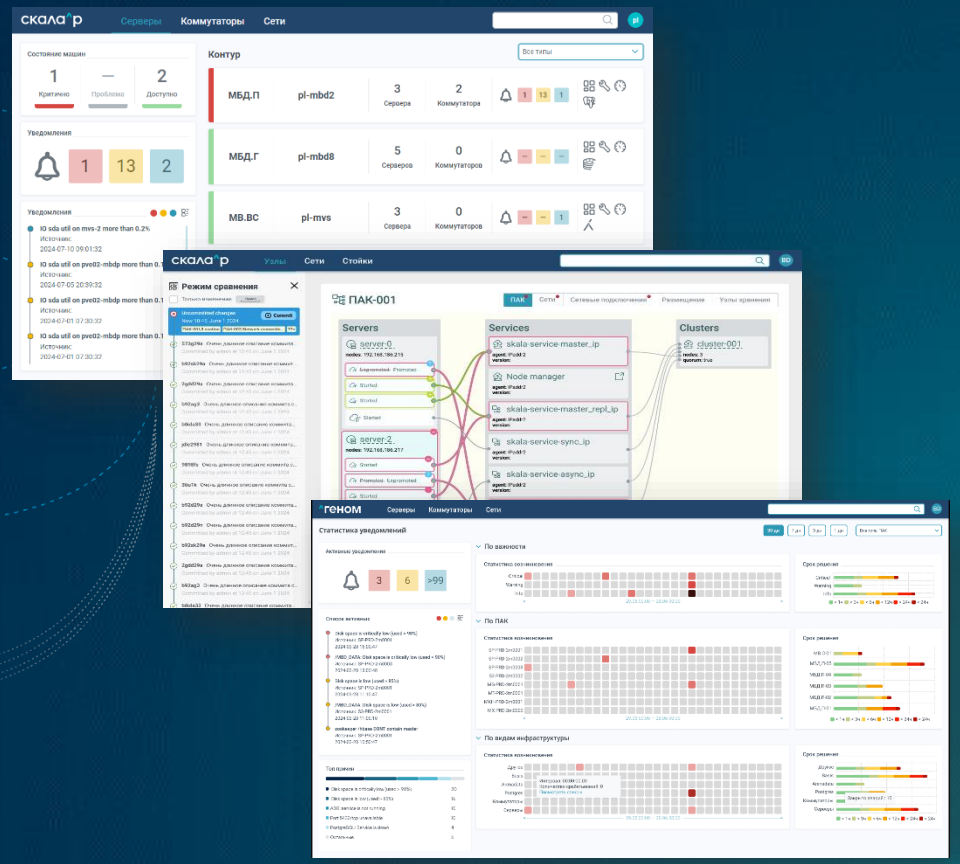
Отраслевые **решения**

Программная платформа Скала^р

Объединения различных доменов управления в единую функциональную графовую CMDB

Комплексное решение для эксплуатации инфраструктуры уровня ЦОД





- Единая точка обзора состояния контура
- Обозримость и удобство управления ЦОД
- Цифровой двойник инфраструктуры
- Контроль изменений быстроменяющихся топологий
- Моделирование изменений в инфраструктуре
- Высокая степень автоматизации
- Построение AI-Copilot для управления ЦОД

Скала^р – Secure by Design

Внутренняя инфраструктура ПАК

- Аппаратное обеспечение (серверы, коммутаторы)
- Служебное ПО
- Прикладное ПО
- СЗИ

API

Существующие системы безопасности

- Системы обеспечения, мониторинга и поддержки ИБ
- SIEM (пример)

- ПАК инфраструктуры Скала МВ.ВРМ, МВ.ДИ, МВ.С
- ПАК хранения данных Скала МБД.П, МХД.О
- ПАК обработки и хранения данных Скала МБД.Х, КХ, Т, Г
- ПАК систем резервного копирования Скала МХД.Р
- ПАК для систем АСУ ТП Скала МВ.П
- ПАК искусственного интеллекта Скала МБД.ИИ

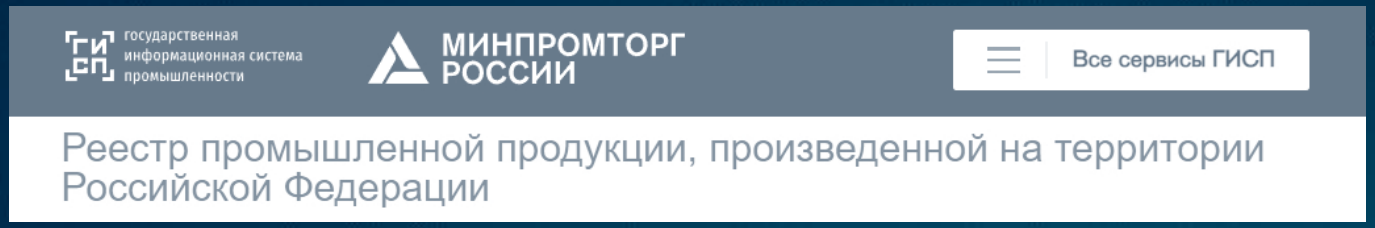
- Использование сертифицированных ОС
- Управление доступом (FAM)
- Антивирус
- Защита виртуальной инфраструктуры
- Использование сертифицированных программных модулей
- Средства доверенной загрузки
- Сканирование на уязвимости на этапе разработки

Результат

- Оптимизация функциональности ПАК с учетом использования средств СЗИ
- Соответствие ПАК аттестационным требованиям

ПАК Скала^р в Реестрах РФ

Машины
Модули
Компоненты



Машины
Модули
Программное обеспечение

Машины
Модули
Программное обеспечение

Соответствуют критериям доверенного ПАК

- Технологическая независимость
- Информационная безопасность
- Функциональная устойчивость

Импортозамещение: сложность выбора
Отсутствие технологического лидерства

Глобальный ИТ-рынок

Сетевая инфраструктура
HUAWEI, CISCO

Хранение данных
HITACHI, DELL, NetApp

Виртуализация
vmware, KVM, Microsoft

Вычислительная инфраструктура
hp, DELL, IBM

СУБД
ORACLE, Microsoft

Операционные системы
Microsoft, Red Hat

Российский ИТ-рынок

Сетевая инфраструктура
ELTEX, QTECH, RDP, Zelax, ЦИФРОВЫЕ РЕШЕНИЯ

Хранение данных
AQUARIUS, AERODISK, RAIDIX

Виртуализация
VeiL, ASTRALINUX, base alt, vmmanager, ROSA, BASIS, NUMA, РУСЭН, ViRT, Горизонт-ВС, ROS ПЛАТФОРМА, КИБЕРПРОТЕКТ

Вычислительная инфраструктура
ГРАВИТОН, ICL, AQUARIUS, OpenYard, kraftway, DELTA COMPUTERS, GAGARIN

СУБД
PostgresPro, AD, tantor, SBER TECH, Yandex, SOQOL, PICODATA, GIS, РЕД БАЗА ДАННЫХ

Операционные системы
base alt, РЕДОС, ROSA, ASTRALINUX, AlterOS, applite, SBER TECH, ОСНОВА, Скейлер

Проблемы отсутствия ИТ-лидеров на российском рынке

- Отсутствие информации и практического подтверждения совместимости продуктов
- Время и ресурсы для подтверждения соответствия заявленной функциональности
- Проблема совместимости с продуктами из разных классов
- Размывание понятия «лидер»: в каждом сегменте существуют десятки на первый взгляд равноценных продуктов

Независимость: варианты реализации

Покомпонентное замещение

- Время на изучение вариантов, тестирование и выбор
- Лавина взаимосвязанных проектов по внедрению
- Сложность синхронизации дорожных карт развития
- Рост сроков внедрения и рисков на стыках

Создание целевой доверенной ИТ-инфраструктуры

- Последовательный перевод систем на целевую доверенную ИТ-инфраструктуру
- Снижение нагрузки с текущей инфраструктуры и необходимости ее масштабирования
- Сокращение сроков внедрения и снижение рисков

Почему ПАК Скала^р ?

- Гарантированно совместимые компоненты
- Отказоустойчивость на уровне архитектуры
- Оптимизация производительности
- Ответственность одного производителя за функционал и показатели назначения
- Решенные вопросы интеграции, эксплуатации, мониторинга, обеспечения ИБ, резервного копирования
- Поддержка и сервис из одного окна
- Серийность и преемственность
- Управляемая дорожная карта развития

Конкурентные преимущества оптимизированных решений Скала^р

Производительность

x2↑

чем решения, использующие сопоставимые аппаратные средства за счет оптимизации ввода-вывода и интерконнекта и за счет разгрузки ЦПУ

x4↑

чем решения в виртуальной среде, использующие сопоставимые аппаратные средства за счет снижения латентности

x4↑

для систем с большим количеством сессий за счет использования специализированных пулеров и балансировщиков

RPO/RTO

x4↓

время выполнения резервного копирования и восстановления за счет специализированного встроенного модуля резервного копирования

x6↓

время полного восстановления узла в случае отказа за счет использования встроенной системы развертывания и цифрового двойника системы

Доступность

Кратное сокращение инцидентов связанных с ошибками эксплуатации и существенное увеличение доступности за счет использования специализированной системы управления ресурсами

** На примере Машины баз данных Скала^р МБД.П*

ПАК — Машины Скала^р — преимущества перед самостоятельными проектами

Высокая отказоустойчивость

За счет специализированной модульной и кластерной архитектуры решений

Высокая производительность

Встречная оптимизация и устранение узких мест по всему стеку применимых технологий

Премиальный сервис

Гарантированно работоспособное решение

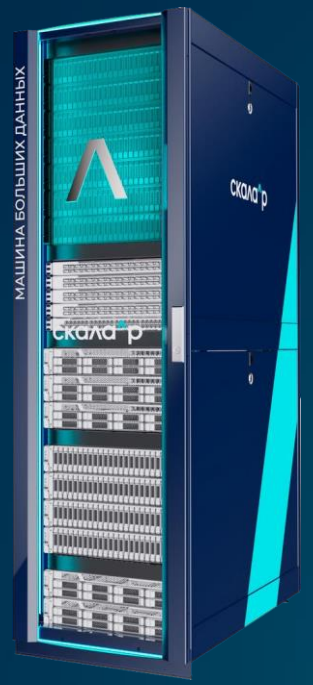
Экономия до 90% на **проектировании и внедрении**

Продукты развиваются с учетом пожеланий Заказчиков

Высокая доступность и катастрофоустойчивость из коробки

Соответствие требованиям **ИБ**

Российское оборудование и ПО



Ускорение до 30% проектов **импортозамещения**

Кратное **сокращение инцидентов,** связанных с ошибками эксплуатации

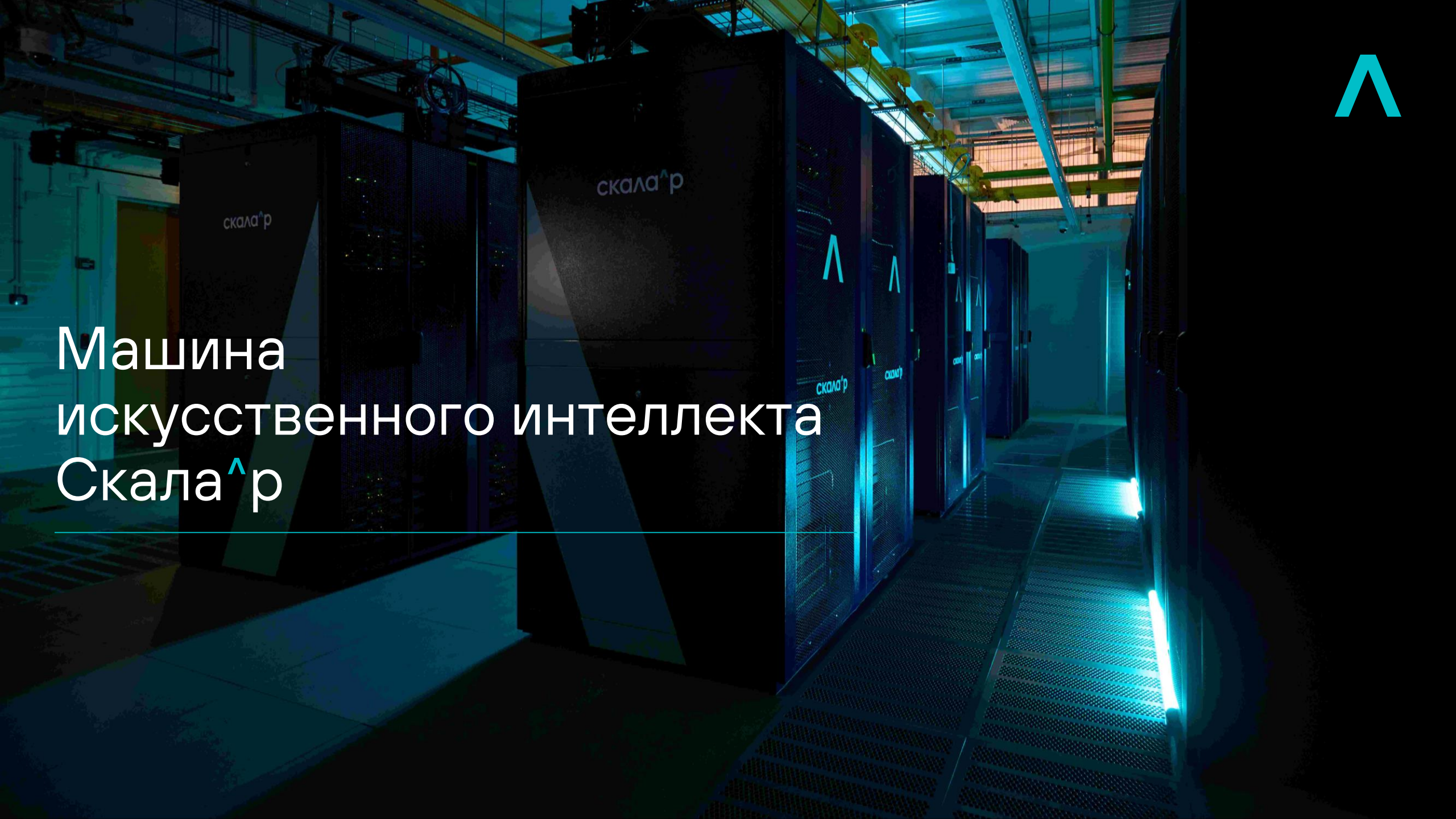
Удобство закупочных процедур для ПАК и Модулей — это **номенклатурные позиции Реестра** РЭП Минпромторга РФ

Соответствие актуальному законодательству по закупкам — **преференции изделиям**

Применение для КИС и ГИС, включая **доверенные ПАК** для КИИ

Прямое взаимодействие с технологическими партнерами по развитию необходимого Заказчикам функционала

ПАК — Программно-аппаратный комплекс и модули платформы — включены в Единый реестр российской радиоэлектронной продукции и реестр Минцифры



Машина искусственного интеллекта Скала^р

Риски и сложности внедрения ИИ

Внедрение ИИ кардинально отличается от внедрения готовых программных продуктов (ERP, CRM, ITSM и т.п.) и от разработки ПО на заказ

- В процессе реализации ИИ-проекта доступен широкий спектр продуктов на рынке ИИ
- Список ИИ продуктов еженедельно меняется и появляются новые как решения так и технологии
- В процессе реализации ИИ-проекта подбираются и апробируются разные ИИ продукты

Для реализации ИИ инициатив нужны разные ИИ-специалисты (DataScience, ML-инженер, DevOps, Аналитик и т.п.)

- Специалистов очень мало на рынке и стоимость высокая
- Высокая конкуренция за ИИ-специалистов
- Долгий срок подготовки и взращивания специалистов даже при наличии наставников
- Зависимость ИИ-проектов от носителя знания или компетенции

ИИ инфраструктура дорогая

- Основная стоимость это графические ускорители и интерконнект
- Требуются специализированные аппаратные решения
- Потенциальные сложности и ограничения при масштабировании ИИ инфраструктуры

Комплексность ИИ решений усложняет организацию ИБ

- Во время интеграции в контур компании есть высокий риск получить дыру в безопасности
- Во время эксплуатации возникают дополнительные риски утечки данных

ПАК ИИ

- ПАК стандартизирует аппаратно-программные решения, гарантируя их совместимость, универсальность, сохранение уровня доверия к ИИ и возможность модернизации
- Готовый набор проверенных моделей и ИИ продуктов, входящих в состав Pandora
- Сокращение срока разработки и внедрения ИИ до 80%

- ПО автоматизации управления ЖЦ Машины ИИ (Pandora) позволяет развивать ИТ-специалистов, что снижает зависимость от рынка труда в сфере ИИ
- IT-специалисты -> MLDevOps -> ML инженеры -> DataScience

- Экономия на дальнейшей масштабируемости за счёт применения HPC технологий и устранении узких мест
- Отработанный план модернизации продукта от тестовой среды до ИИ-фабрики
- Обеспечена вариативность между NVIDIA, азиатскими вендорами и другими
- Оптимизация требуемых ресурсов от задачи к задачи

- Обеспечение безопасности ИИ сервисов за счёт подхода TRISM/secure by design и каталога доверенных контейнеров, готовых к запуску на on-premise инфраструктуре, в закрытом контуре
- ПАК может функционировать в изолированном контуре, предотвращая утечки данных
- Использование доверенных ИИ компонентов из его каталога и листа совместимости, минимизирует риски компрометации конечного продукта

₽/Token

₽/TOPS

₽/FLOPS

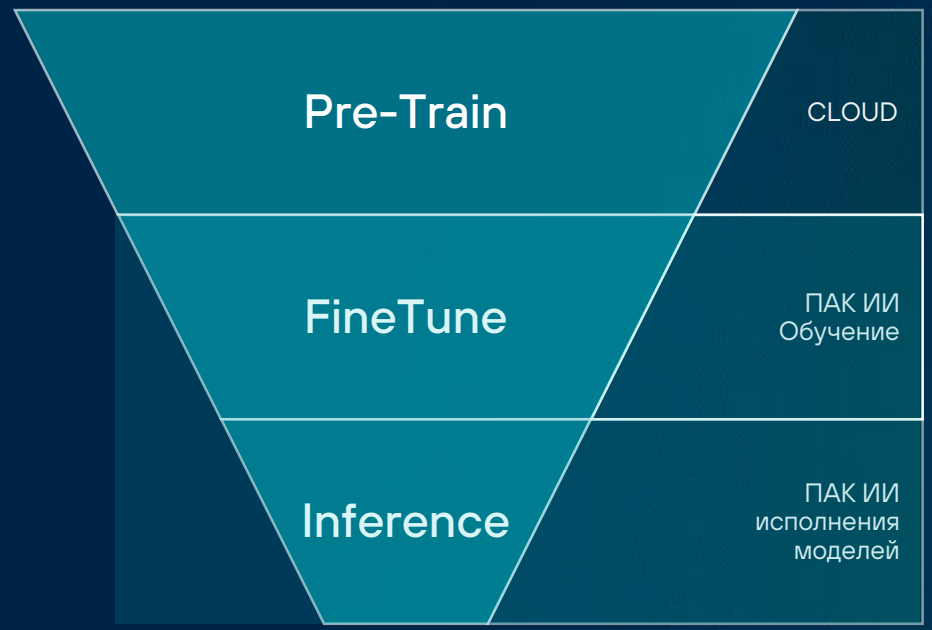
Как итог улучшения показателя рубль за единицу вычисления

Импортозамещение: Соответствие требованиям регуляторов и уменьшение санкционных рисков

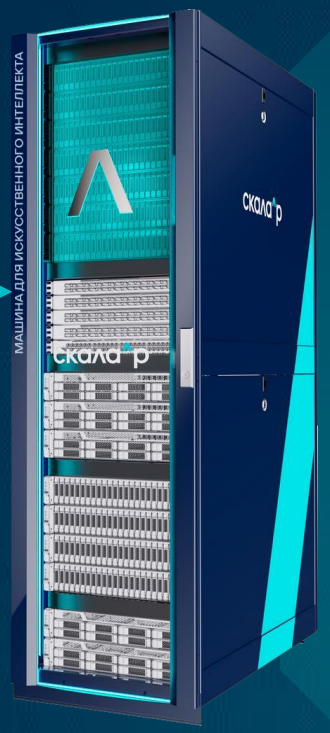
Машина искусственного интеллекта Скала^р

ИИ решения

Cotype · GigaChat · ValueAI · Llama · Сайбокс · DeepSeek · YandexGPT 5 · Смарт Платформа · Другие ИИ решения



ПАК позиционируется для обеспечения on-premise инфраструктуры для обучения и исполнения ИИ



Аналог
Huawei Atlas 900 Pod / NVIDIA DGX SuperPOD

Каталог контейнеров — Подготовленные ИИ решения · Профилированные контейнерные среды

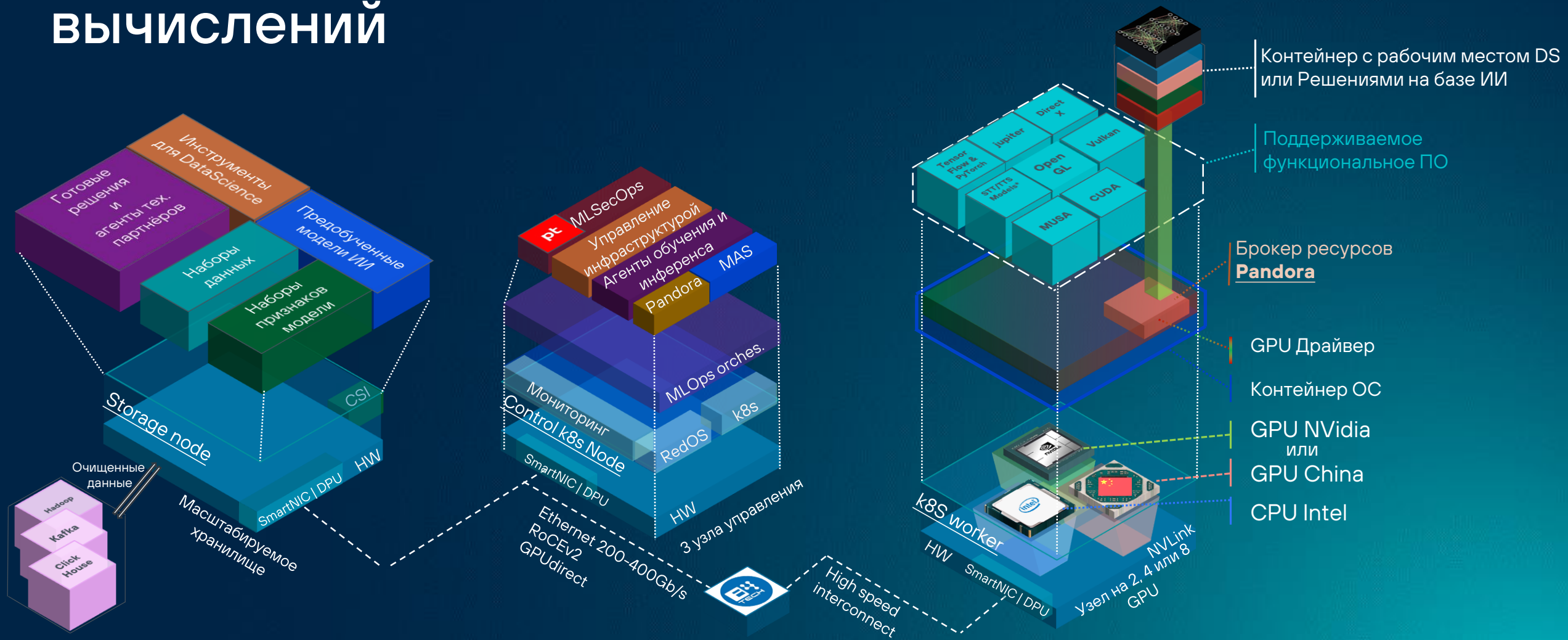
ML/LLM пространство — Обучение моделей · Разработки ИИ решений · Исполнение моделей ИИ

Контейнерная инфраструктура

Аппаратное обеспечение: Вычисления · Хранение · Сеть передачи данных

Информационная безопасность

В основе новой Машины ИИ Скала^р — передовые технологии распределенных вычислений



Размеры Машины Скала^р МИИ

Тестирование / Демо	Разработка / MVP	Продакшн (корпоративный)	AI Factory (полный цикл)
Сеть: Сеть опционально	Сеть: Сеть доступа; Сеть управления	Сеть: Сеть доступа/хранения; Сеть интерконнекта; Сеть управления	Сеть: Сеть доступа; Сеть интерконнекта; Сеть хранения; Сеть управления
Вычисления: Сервер на 4-8 GPU + NVLink или Сервер на 2 GPU	Вычисления: Сервер на 4-8 GPU + NVLink и/или Сервер на 2 GPU	Вычисления: 2 сервера на 4-8 GPU + NVLink и/или 2 сервера на 2 GPU	Вычисления: Сервер (-ы) на 4-8 GPU + NVLink и/или Сервер (-ы) на 2 GPU
Управление: Управление опционально	Управление: Сервер управления	Управление: 2 сервера управления	Управление: 3 сервера управления
Хранение: Хранение локально	Хранение: Хранение 1) Локально 2) Внешний S3	Хранение: Хранение 1) Локально 2) Внешний S3	Хранение: Масштабируемое хранилище объектное
S	M	L	XL

Параметры продуктовой линейки

Функция / Характеристика	S (Small)	M (Medium)	L (Large)	XL (Extra Large)
Количество узлов	1 (Модуль инференса)	2+ (Минимальный кластер)	4+ (Отказоустойчивый кластер)	10+ (Масштабируемый кластер)
Поддержка инференса	✅ Да (локальный)	✅ Да (кластерный*)	✅ Да (оптимизированный)	✅ Да (масштабируемый)
Поддержка GPU/TPU	⚠ 1 GPU (опционально)	✅ Да (несколько GPU)	✅ Да (кластер GPU)	✅ Да (оптимизированные фермы)
Мониторинг и метрики	⚠ Базовые метрики	✅ Prometheus + Grafana	✅ Расширенная аналитика	✅ AI-аналитика + предсказания
Kubernetes (k8s) Management	❌ Нет	✅ Да (базовое управление)	✅ Да (продвинутое управление)	✅ Да (полный контроль + мониторинг)
Отказоустойчивость	❌ Нет	⚠ Частично	✅ Да (автовосстановление)	✅ Да (высокая доступность)
Создание ИИ-агентов	❌ Нет	⚠ Базовые сценарии	✅ Да (сложные агенты)	✅ Да (автономные агенты)
Масштабируемость	❌ Нет	⚠ Ручное масштабирование	✅ Да (автоматическое)	✅ Да (гибкое + балансировка)
ИИ-ассистенты	❌ Нет	❌ Нет	⚠ Простые интеграции	✅ Да (многомодальные ассистенты)
Обучение моделей	❌ Нет	❌ Нет	⚠ Ограничено	✅ Да (распределённое обучение)
Целевой сценарий	Тестирование / Демо	Разработка / MVP	Продакшн (корпоративный)	AI Factory (полный цикл)

** С добавлением модуля управления, можно кластеризировать модули инференса*

Примеры исполняемых задач

с применением GPU платформы с NVLink

- YandexGPT (LLM)
- SciBox (Интсрумент)
- RAG (Инструмент)
- Cotype (LLM)
- DeepSeek (LLM)
- ValueAI (Инструмент)
- Llama (LLM)
- GigaChat (LLM)
- Другие ИИ решения

- Анализ транзакционной сети (миллионы узлов) в режиме реального времени для выявления сложных схем мошенничества.
- Анализ кредитной истории + текстовых данных (договоры, переписка).
- Детекция аномалий в потоке транзакций (~100K TPS)
- Распознавание и верификация голоса в колл-центрах банка
- Парсинг договоров, регламентов, сканов документов и выявление рисков
- Обработка тысяч источников для прогноза волатильности рынка
- Автоматическое формирование отчетов по регуляторике на основе внутренних данных

Примеры исполняемых задач

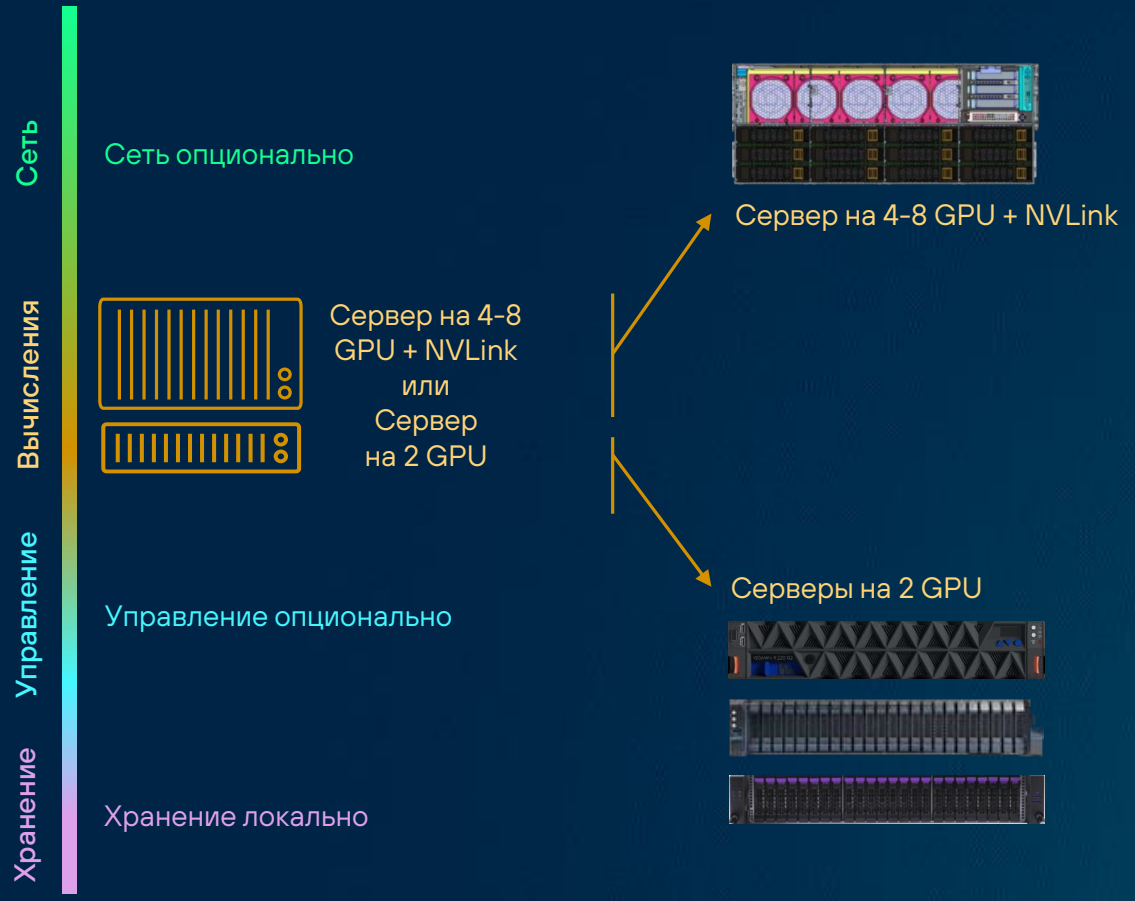
с применением типовых серверов 2RU

- YandexGPT (LLM)
- SciBox (Интсрумент)
- RAG (Инструмент)
- Cotype (LLM)
- DeepSeek (LLM)
- ValueAI (Инструмент)
- Llama (LLM)
- GigaChat (LLM)
- Другие ИИ решения

- Извлечение данных из документов
- Прогнозирование оттока клиентов
- Классический кредитный скоринг с фичами из транзакций.
- Выявление подозрительных транзакций (но не в реальном времени)
- Анализ клиентских профилей
- Ответы на типовые вопросы клиентов (без сложного RAG)
- Автоматическое категоризирование расходов. Разметка транзакций
- Проверка паспортов, договоров через компьютерное зрение

Инференс-узлы

тестирование/демо

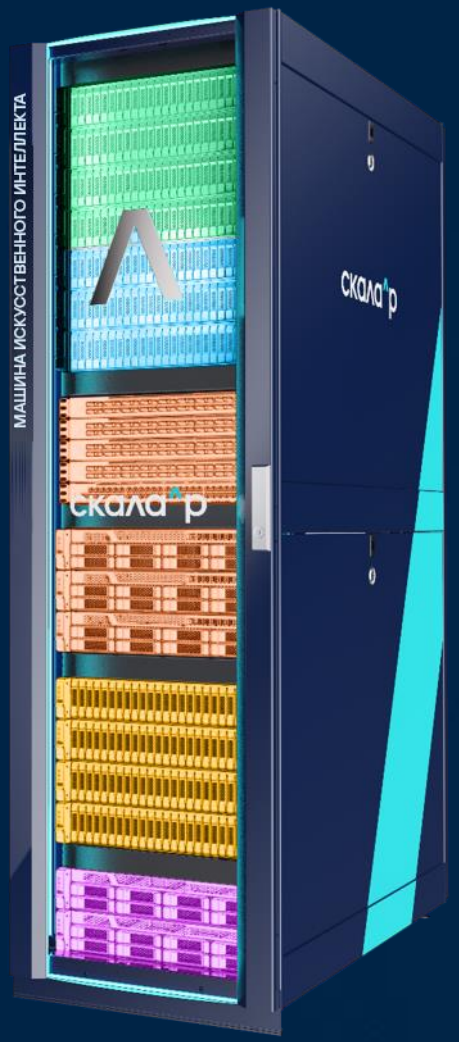


Юниты	4 RU
Процессоры	2x Intel Xeon 4/5 Gen
RAM	DDR5
PCIe слоты	Передняя панель: поддерживает максимум один слот PCIe 5.0 Задняя панель: поддерживает максимум 10 слотов PCIe 5.0 и 8 видеокарт двойной ширины
Электропитание	~3,6КВт

Юниты	2 RU
Процессоры	2x Intel Xeon 4/5 Gen
RAM	DDR5
PCIe слоты	До четырёх PCIe 5.0 x16 и до трёх PCIe 5.0 x8
Электропитание	~1,3КВт

Используемые GPU NVIDIA	NVIDIA H100	NVIDIA H200	NVIDIA A100	NVIDIA L40s	NVIDIA T4/L4
Используемые GPU Азия	16GB GDDR	32GB GDDR	48 GDDR	Аналог NVLink органичено, PCIe 4.0 и PCIe 5.0	

Машина Скала^р МИИ — Модули



Модуль полезной нагрузки Машины МИИ

- Bare metal узлы, выступающие в качестве Worker нод кластера Deckhouse Kubernetes Platform. Количество этих узлов можно варьировать от 3 до 16 (в некоторых случаях возможна конфигурация от 1 узла).
- Вычислительные мощности узла — 64 физических ядра CPU, от 128ГБ до 4ТБ ОЗУ при оптимальной конфигурации памяти.
- От 1 до 8 GPU типа H100 в один узел
- Диски в этих узлах (от 4 штук в каждом узле в базовой конфигурации с возможностью масштабирования до 16 дисков на узел) можно использовать для организации хранения данных контейнеров, на сегодня это опции local path provisioner и SDS local volume в терминологии Deckhouse Kubernetes Platform.

Базовый модуль

Коммутационный модуль Машины МИИ

- Два коммутатора 100GbE или 400GbE на 32 порта(каждый) в отказоустойчивой конфигурации для сети интерконнекта Машины.
- Два коммутатора от 25GbE по 48 портов в отказоустойчивой конфигурации для организации доступа к сервисам Машины МИИ из сети заказчика.
- Два коммутатора от 25GbE на 48 портов(каждый) для организации сети хранения данных Машины.
- Два коммутатора 1GbE на 48 портов (каждый) для организации управляющей сети (out-of-band управление и in-band управление).

Модуль управления Машиной МИИ

- Три сервера для размещения управляющих компонент Машины – управляющих и служебных узлов Deckhouse Kubernetes Platform, сервисов Скала^Р.
- Диски в этих узлах (по 4 штуки в каждом узле в базовой конфигурации с возможностью масштабирования до 16 дисков на узел) можно использовать для организации различных вариантов хранилищ.

Модуль хранения Машины МИИ

- Подключаемый к кластеру DKP Машины контейнерной инфраструктуры посредством CSI драйвера.
- Поддержка распределенных вычислений
- Поддерживает многопоточную загрузку/выгрузку (например, через s5cmd, rclone)

Машина Скала^р МИИ XL — компоненты

Коммутационный модуль Машины МИИ

- Сеть управления 1GbE (2 x B4Com CS2148-4D)
- Сеть доступа 25GbE (2 x B4Com CS4148Q-8U)
- Сеть хранения данных 25GbE (2 x B4Com CS4148Q-8U)
- Сеть интерконнекта 100GbE или 400GbE (2 x B4Com CS4132U или 2 x B4COM CS6132F)

Модуль управления Машиной МИИ (3 физических сервера)

Deckhouse Kubernetes Platform 1.67 CSE control plane и системные ВМ

Мониторинг 1	DKP сервисы 1	AI Task manager
K8s управление 1	DKP балансировщик 1	HW MGMT

Ред ОС 7.3 ФСТЭК (qemu, kvm, libvirt)
Сервер модуля управления 1

Мониторинг 2	DKP сервисы 2	AI Task manager
K8s управление 2	DKP балансировщик 2	HW MGMT

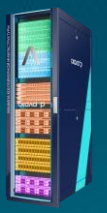
Ред ОС 7.3 ФСТЭК (qemu, kvm, libvirt)
Сервер модуля управления 2

Мониторинг 3	DKP сервисы 3	AI Task manager
K8s управление 3	DKP балансировщик 3	HW MGMT

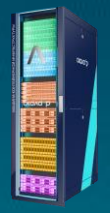
Ред ОС 7.3 ФСТЭК (qemu, kvm, libvirt)
Сервер модуля управления 3

Модуль хранения Машины МИИ

Система хранения данных


МБД.8

и/или


МХД.О

Модуль полезной нагрузки Машины МИИ (от 3 до 16 физических серверов, до 8GPU на узел)

Deckhouse Kubernetes Platform 1.67 CSE worker nodes

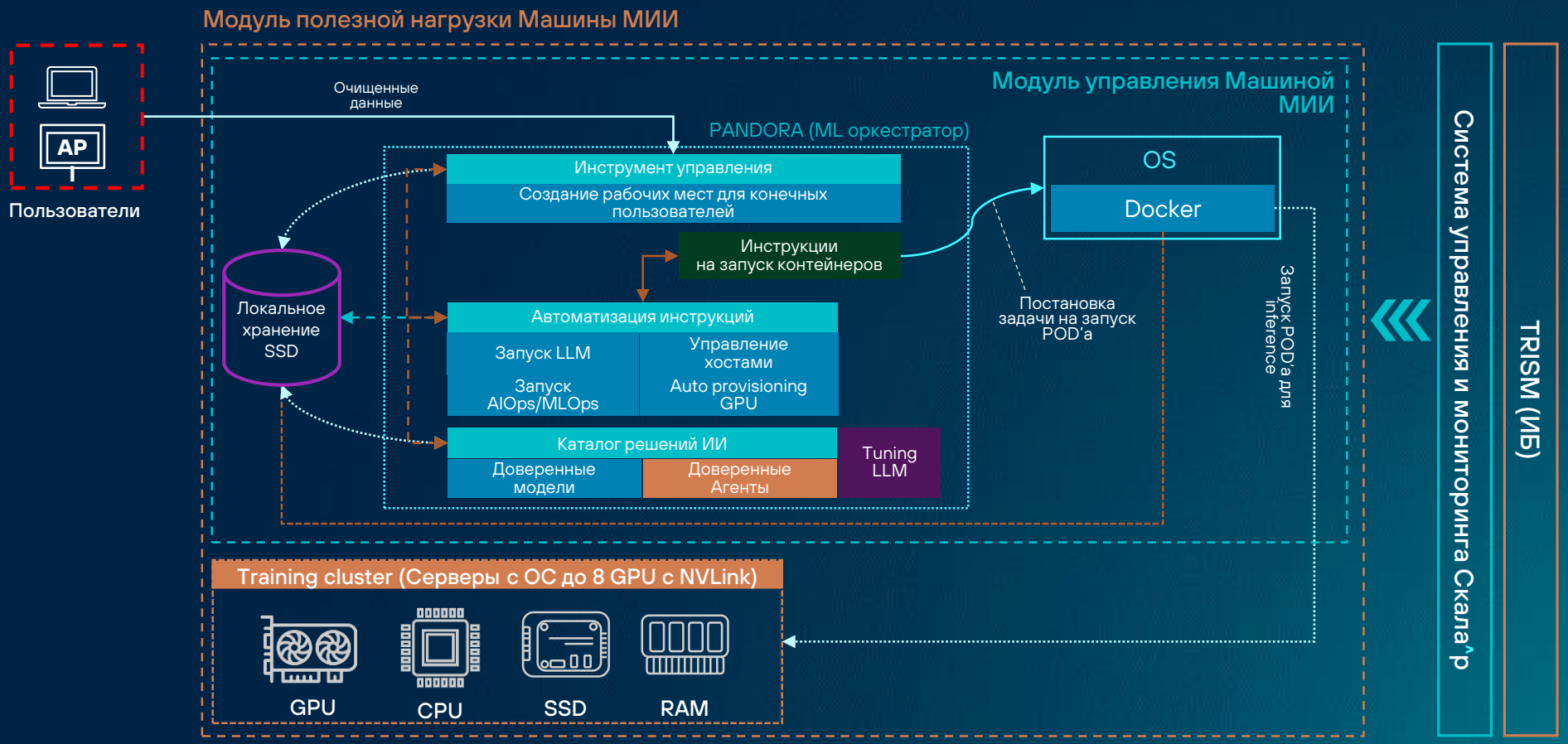
ОС Ред ОС 7.3 ФСТЭК bare metal
- Сервер модуля полезной нагрузки 1
- Сервер модуля полезной нагрузки 2
- Сервер модуля полезной нагрузки 3
- Сервер модуля полезной нагрузки 4

ОС Ред ОС 7.3 ФСТЭК bare metal
- Сервер модуля полезной нагрузки 5
- Сервер модуля полезной нагрузки 6
- Сервер модуля полезной нагрузки 7
- Сервер модуля полезной нагрузки 8

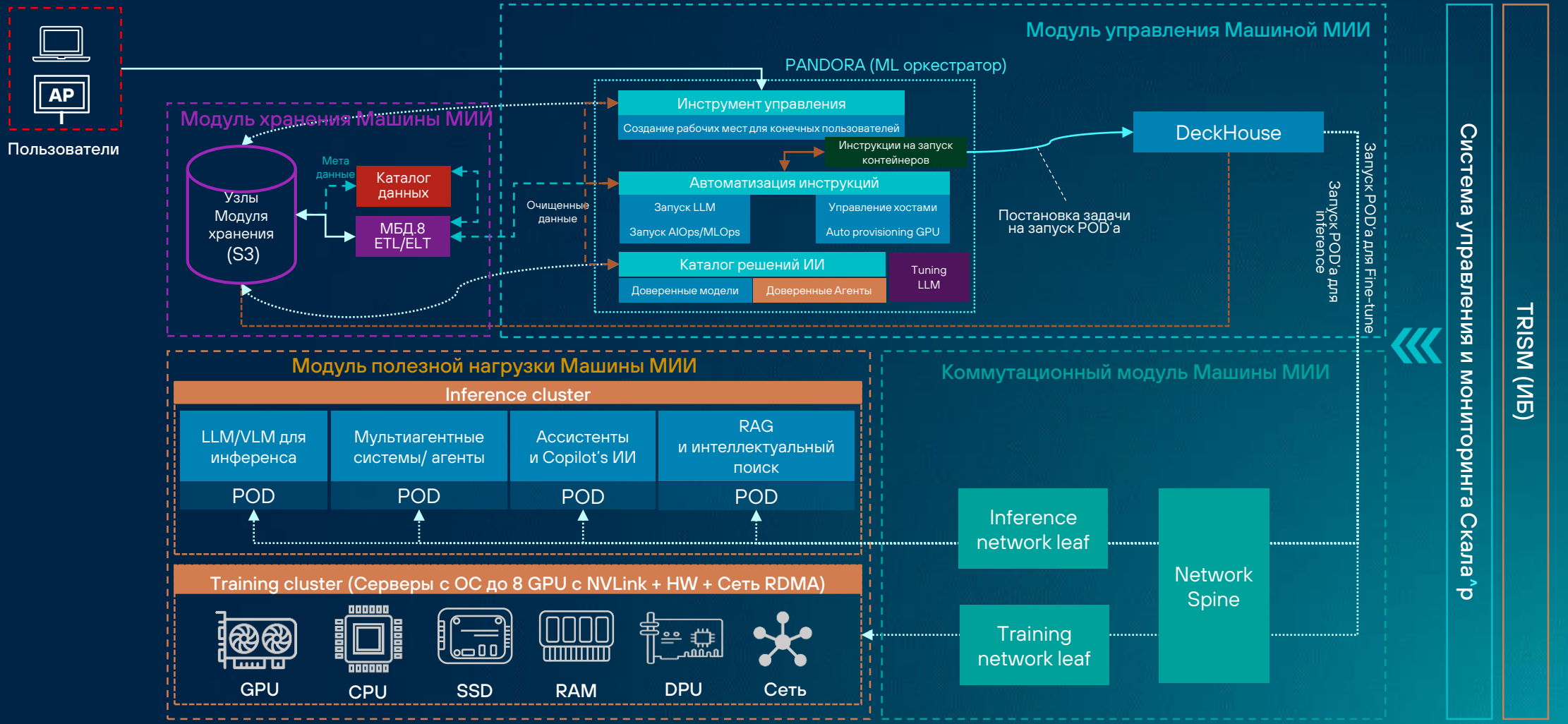
ОС Ред ОС 7.3 ФСТЭК bare metal
- Сервер модуля полезной нагрузки 9
- Сервер модуля полезной нагрузки 10
- Сервер модуля полезной нагрузки 11
- Сервер модуля полезной нагрузки 12

ОС Ред ОС 7.3 ФСТЭК bare metal
- Сервер модуля полезной нагрузки 13
- Сервер модуля полезной нагрузки 14
- Сервер модуля полезной нагрузки 15
- Сервер модуля полезной нагрузки 16

Машина Скала^р МИИ «S» — логическая схема

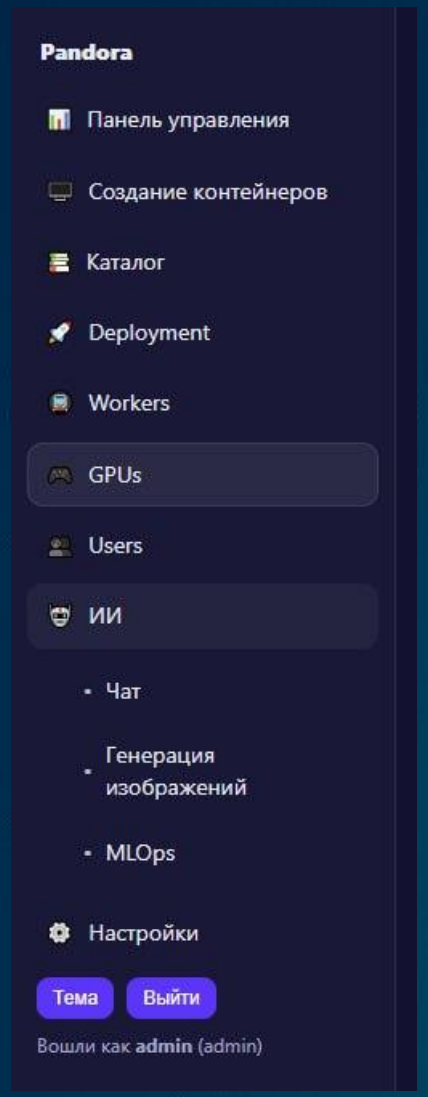


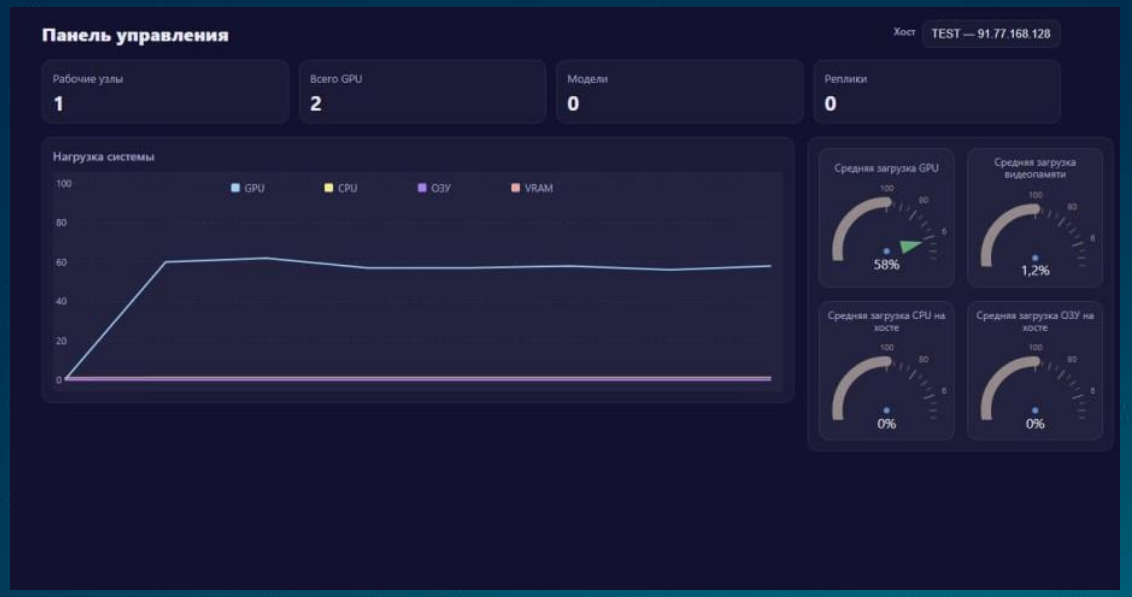
Машина Скала^р МИИ «XL » — логическая схема

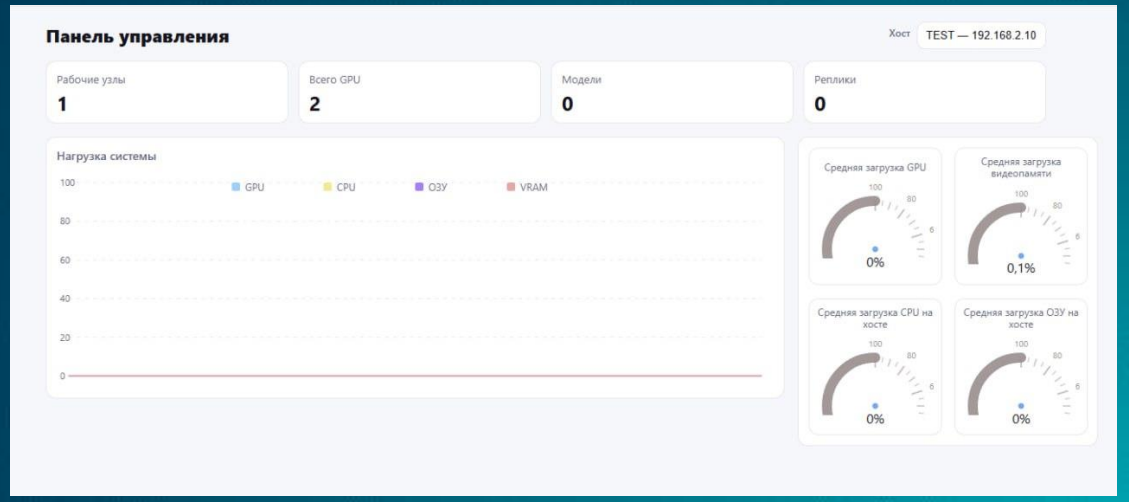


PANDORA (главная страница с мониторингом)

- Мониторинг ресурсов (GPU, CPU, RAM, vRAM)
- Можно увидеть количество хостов в кластере и кол-во GPU
- Кол-во запущенных ПОДов
- Кол-во используемых токенов (всего)
- Кол-во одновременных сессий в сторону модели/моделей (RPS)
- Кол-во ассистентов/агентов ИИ
- Тёмная и белая темы везде

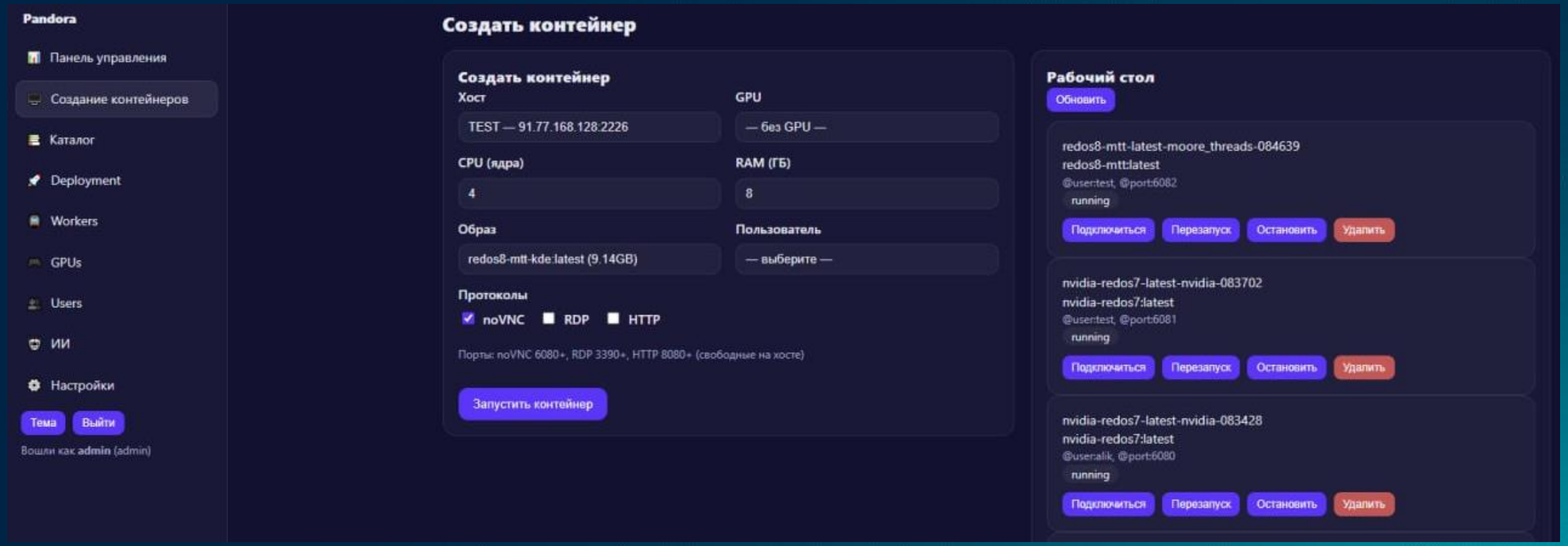






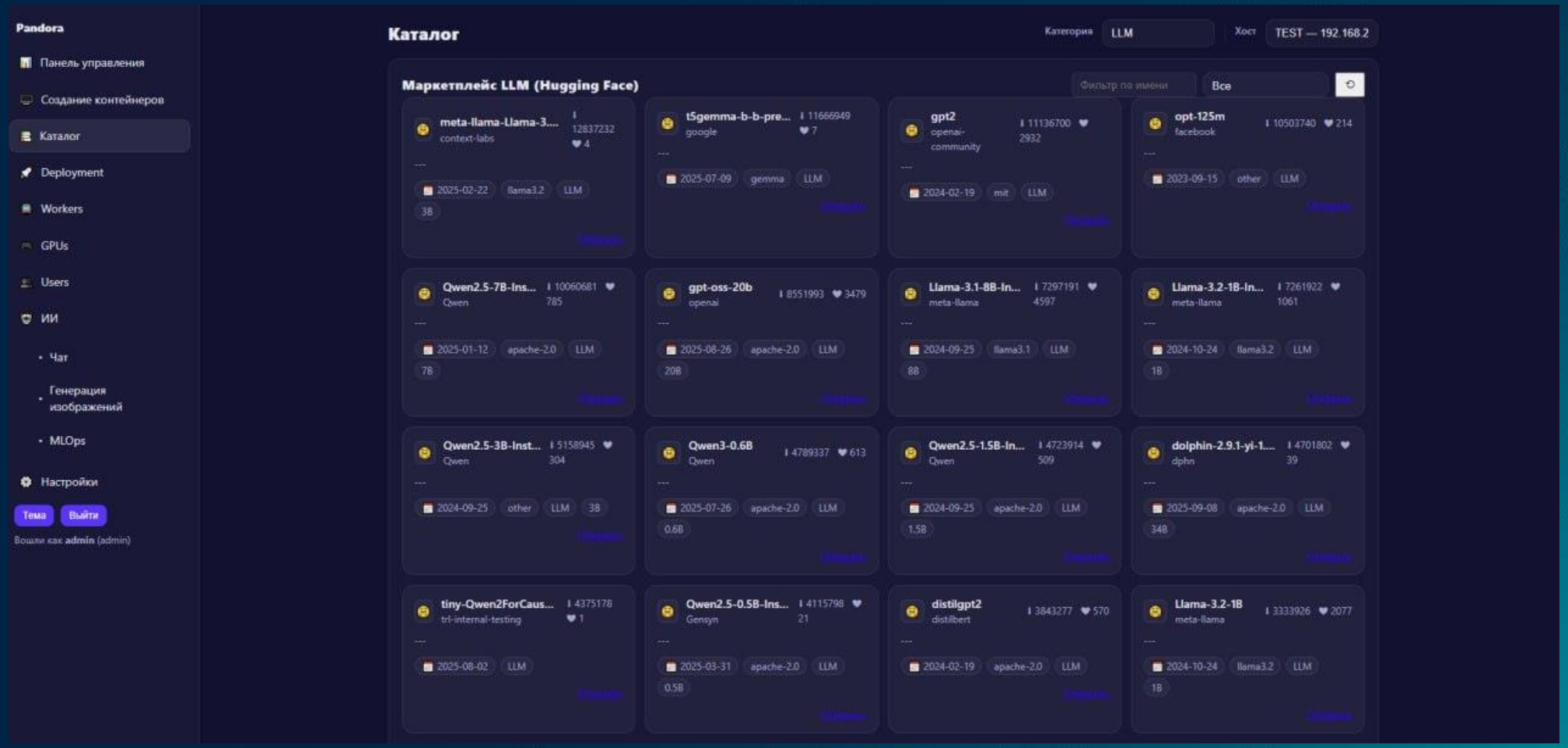
PANDORA

Создание рабочих мест с GPU ресурсами для конечных пользователей
(разработчики, ML/DS-инженеры, бухгалтеры, юристы, сервисная поддержка и т.д.)



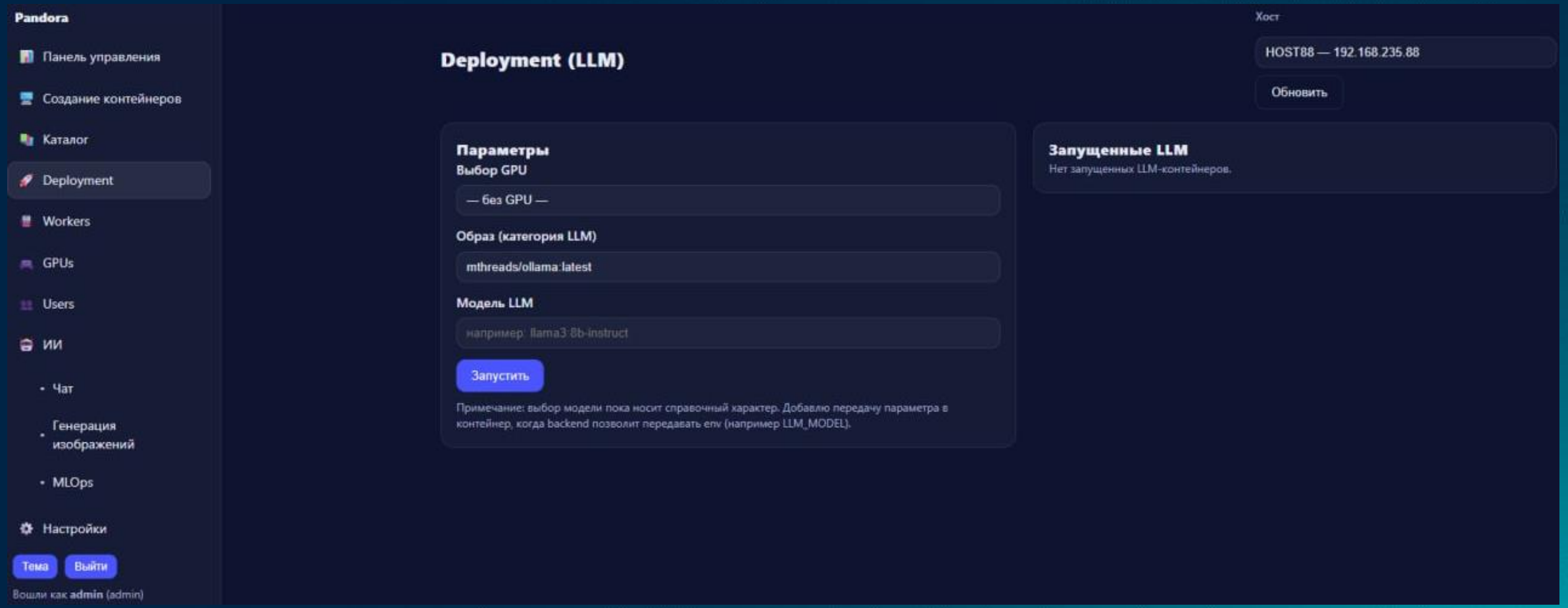
PANDORA

Каталог контейнеров и партнёрских решений, готовых к быстрому запуску на ПАК ИИ. Так же может пополняться решениями и разработками заказчика.



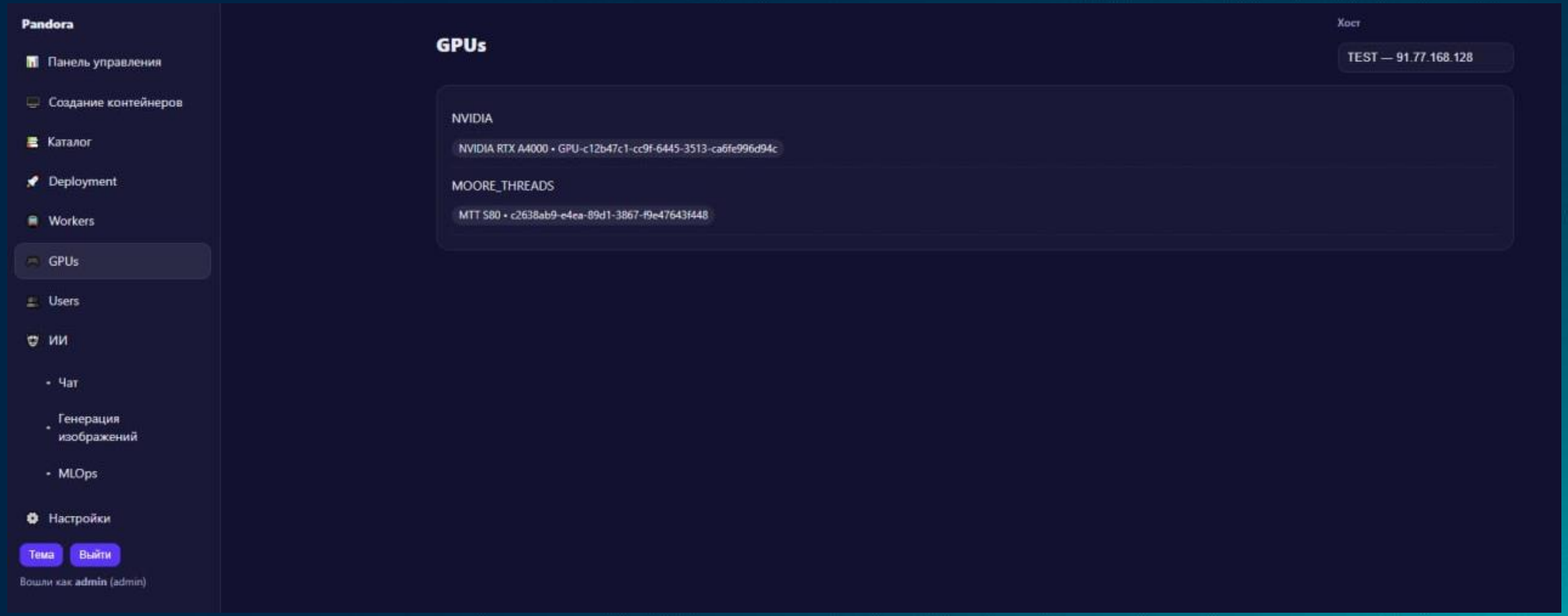
PANDORA

Отдельный интерфейс для создания и запуска контейнера с LLM моделью на ПАК ИИ



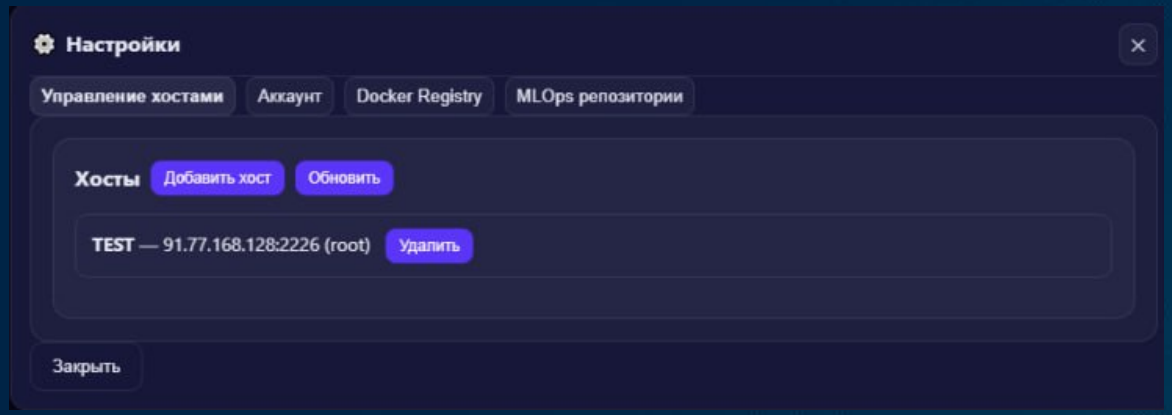
PANDORA

Интерфейс с представлением GPU на ПАК ИИ для хоста/блока или ПАК ИИ.

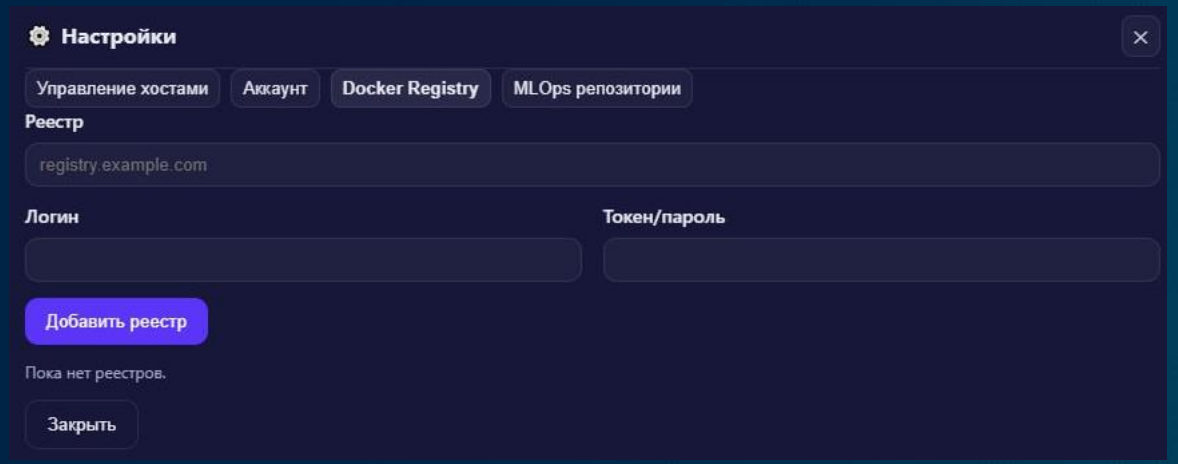


PANDORA

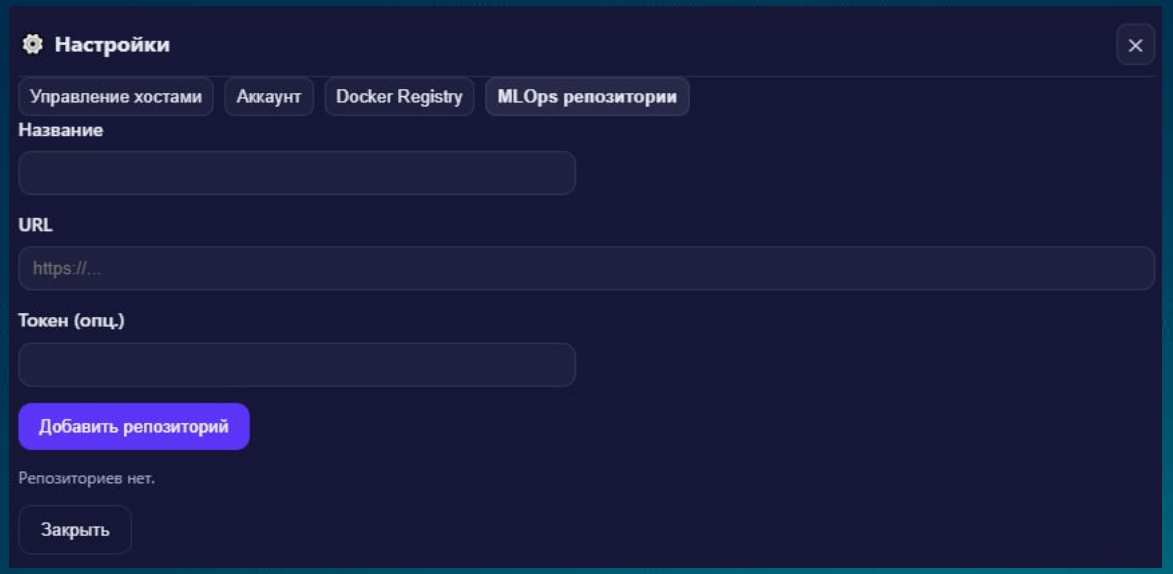
Модальное окно управления хостом (-ами)



Модальное окно добавления репозиториев docker

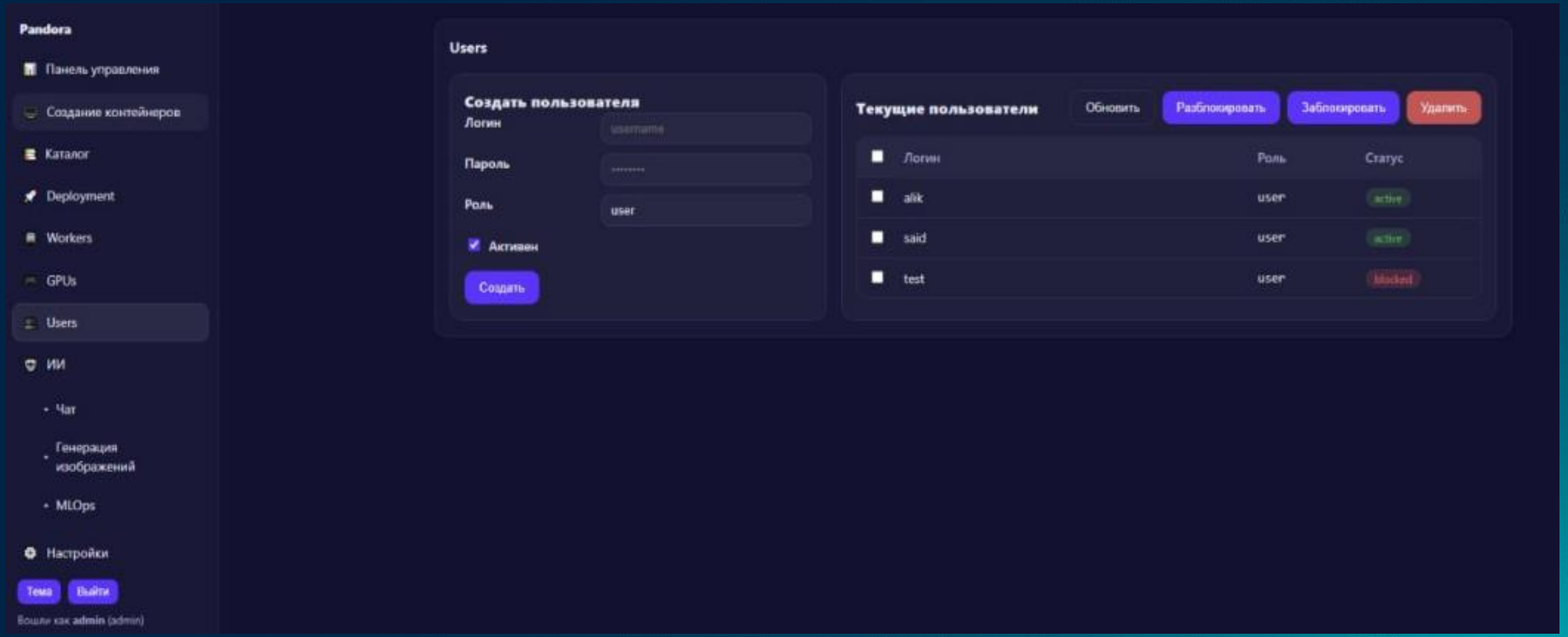


Модальное окно добавления репозиториев MLOps для LLM



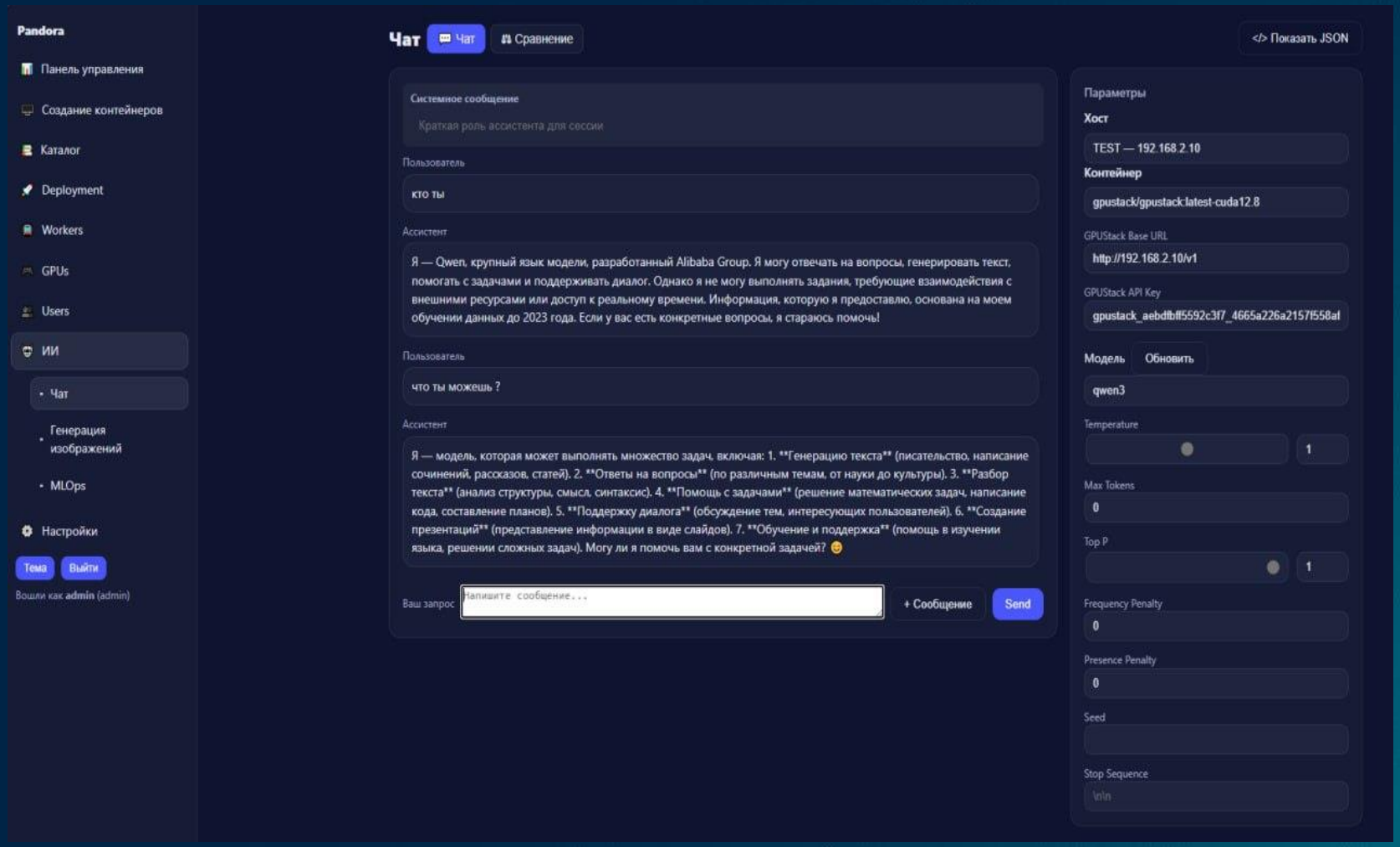
PANDORA

Интерфейс управления правами доступа



PANDORA

Интерфейс работы с развёрнутой LLM на хосте/модуле/ПАК ИИ , а так же Fine-tune модели



Метрики по ПАКу ИИ

Производительность

>= 6 Pflops

на один рабочий узел (TF32)

>= 400 Tflpos

на один рабочий узел (FP32)

Масштабирование сети с применением RDMA увеличивает производительность и уменьшает задержки более чем в 2 раза, что критично для моделей ИИ, работающих на распределённом кластере со сложной топологией

Максимальный размер LLM

до 188B параметров

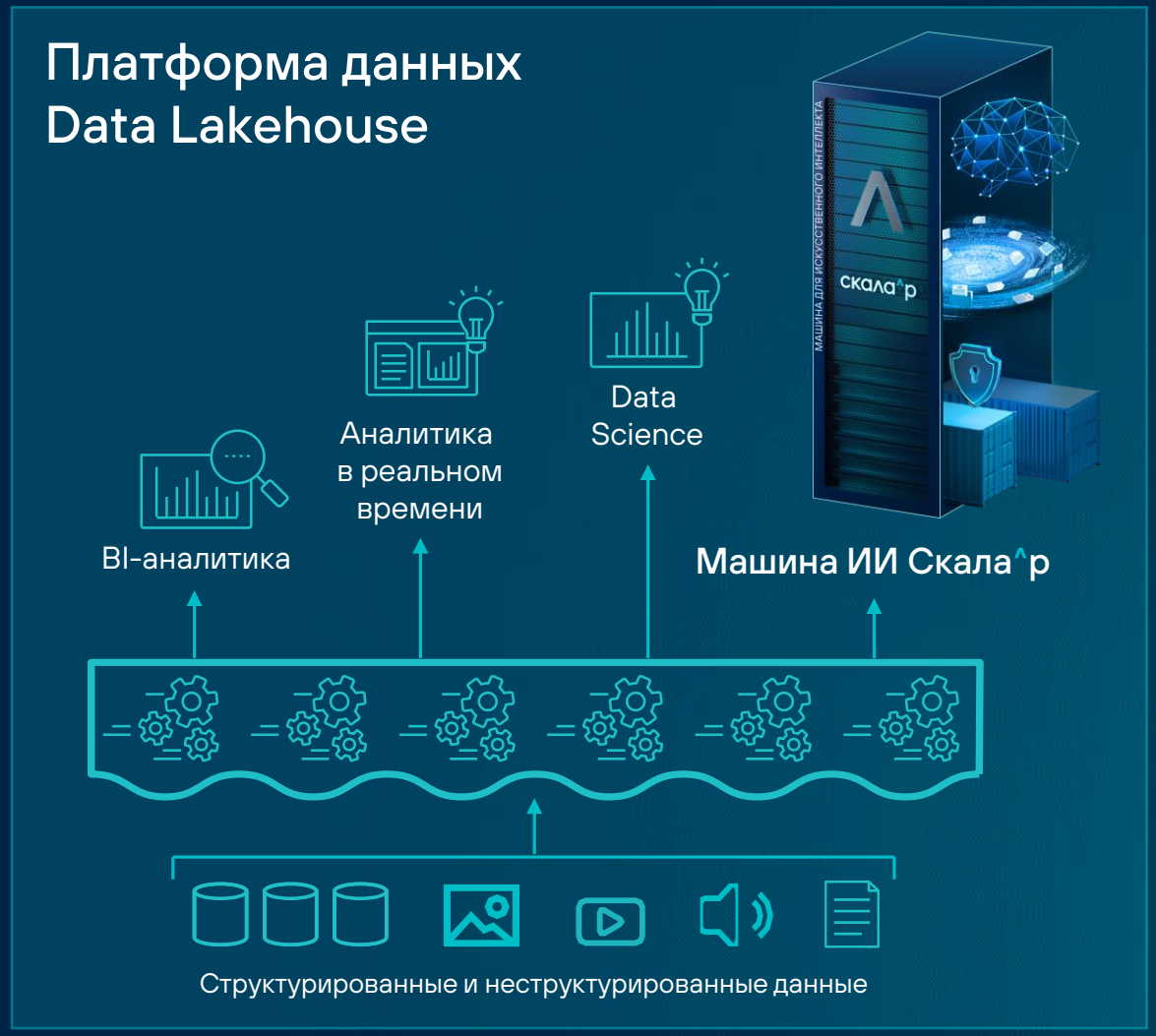
в один рабочий узел без квантования

Использование NVLink для LLM увеличивает TPS примерно

в 2-5 раз

в зависимости от количества пользователей и типа запросов

Экосистема данных Lakehouse + ML/AI



Единое управление данными

объединяет структурированные и неструктурированные данные для обучения моделей ИИ, поддерживая разнообразные рабочие нагрузки (например, BI, ML, генеративный ИИ)

Расширенные возможности для ИИ

применение инновационных форматов с поддержкой транзакционности и версионирования, гарантии надёжности данных для ИИ

Масштабируемость и производительность

оптимизировано для крупномасштабного использования ИИ с инструментами, поддерживающими аналитику в реальном времени

Управление и безопасность

качество данных и соответствие требованиям для приложений ИИ

Интеграция генеративного ИИ

обеспечивает инновационные варианты использования, такие как агенты и системы рекомендаций

Преимущества Машины ИИ Скала^р

Платформенные решения позволяют сократить
- в 15 раз время подготовки среды разработки
- в 5 раз время работы дата-инженеров и дата-аналитиков*

Реализация каталога ИИ решений от валидированных партнеров на базе ПАК ИИ

Надежная мультивендорная Enterprise-инфраструктура с оптимальной конфигурацией и стабильным программно-аппаратным стеком на основе проведенных тестов и лучших практик.

Соответствие требованиям соблюдения принципов отказоустойчивости, масштабируемости на уровне архитектуры для использования в критичных и высоконагруженных корпоративных и государственных информационных системах

Исключение инцидентов на стыке технологий и высококвалифицированная поддержка Скала^р

Расширение возможностей как вертикального, так и горизонтального масштабирования

Предсказуемые характеристики, метрики функционирования платформенных решений

Управление жизненным циклом корпоративных ИИ решений

Поддержка регуляторных требований, отраслевых стандартов

Увеличение производительности*
- в 3 раза при обучении ML моделей
- в 4 раза обученных ИИ моделей

Безопасное использование популярных языковых моделей LLM в закрытом контуре

* Показатели могут варьироваться в зависимости от задачи

Примеры использования ИИ для корпоративных задач*

1

Совершенствование процессов технической поддержки продуктов компании IT_ONE

Автономная система для классификации, маршрутизации поступающих обращений клиентов по разным каналам связи на корректную линию технической поддержки.

Построена на основе обработки естественного языка с применением адаптированных языковых моделей LLM.

2

Повышение эффективности клиентского сервиса

Чат-бот технической поддержки клиентов для информирования, ответов на общие вопросы, уточнения дополнительной информации.

Построен на основе технологии обработки естественного языка и дообученных языковых моделях LLM.

3

Совершенствование внутренних процессов по повседневной работе сотрудников

Расшифровка аудиозаписей встреч с суммаризацией итогов, определения решений и поручений по аудиозаписи: на основе обработки естественного языка, транскрибация, применение адаптированных языковых моделей LLM.

4

Создание единого связанного пространства данных из разнородной информации документов ограниченного доступа, приходящих в ответ на запросы контролирующих органов государственной власти федерального уровня

Автономное (on-premise) ИИ-решение на основе LLM, в формате ПАК для автоматического извлечения данных из неструктурированных документов и автоматического формирования фабулы документа с гибкой настройкой правил извлечения данных.

5

Повышение эффективности разработки и тестирования программных продуктов компании

Чат-ботов для разработчиков и тестировщиков, с поддержкой используемых языков программирования с учетом кодовой базы клиентских продуктов (ПО) во внутреннем контуре компании.

Создание изолированной ИТ-инфраструктуры для эксплуатации результатов инициатив ИИ.

6

Повышение эффективности процессов управления проектами компании

Интеллектуальный помощник (чат-бот), повышающий эффективность повседневной работы руководителей проектов с внутренней документацией, базой знаний и регламентами компании, хранящимися в разнородных внутренних корпоративных сервисах компании.

Построен на основе адаптированных языковых моделей LLM, интеллектуального алгоритма для контекстного поиска, агрегации данных и предоставления структурированных ответов через интуитивный интерфейс чата.

7

Формирование у сотрудников компетенций, позволяющих использовать доверенные технологии ИИ

Средства обучения сотрудников промпт-инжинирингу и мотивации использования ИИ на основе.

Построены на больших фундаментальных языковых моделях (облачных) для выполнения текущих задач.

8

Совершенствование процессов подбора сотрудников

Система скрининга соискателей на соответствие требованиям позиции (вакансии).

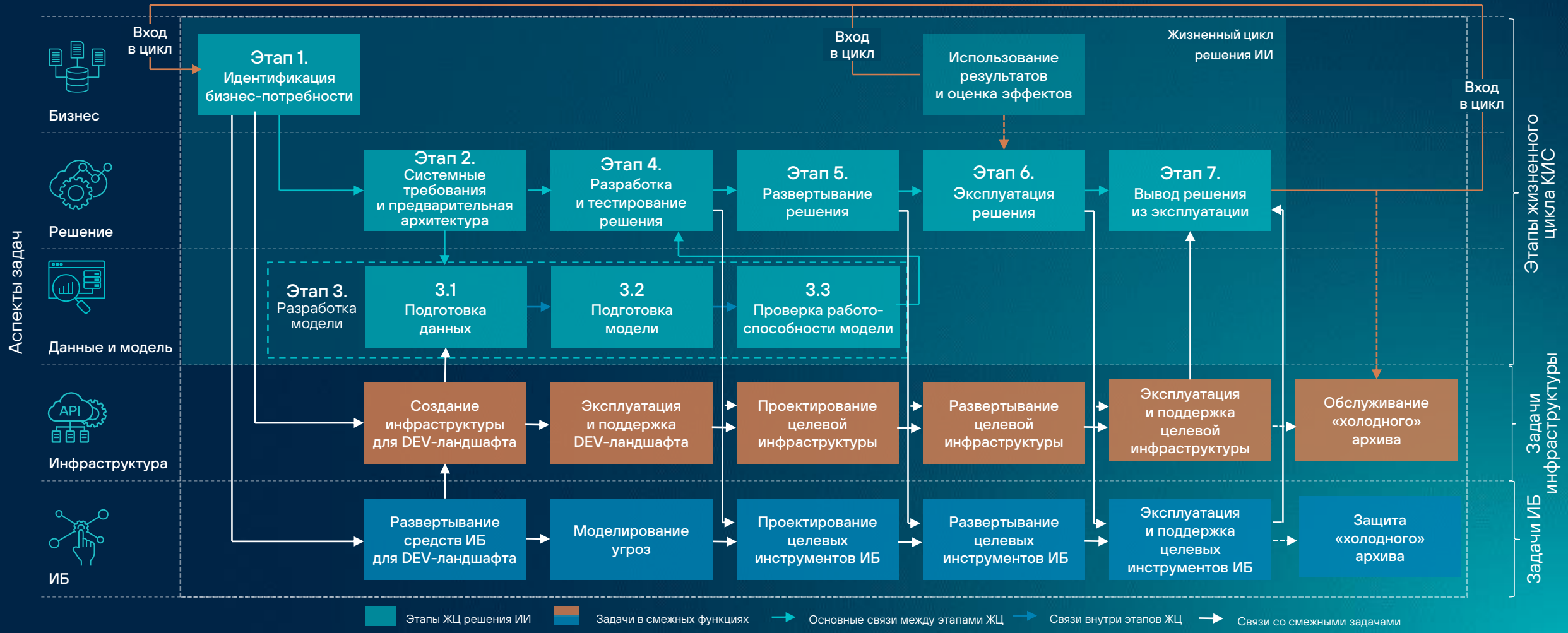
Построена на основе технологий NLP и применения адаптированных языковых моделей LLM.

* Типовые задачи для инфраструктуры Машины ИИ Скала^р

Жизненный цикл КИС с ИИ

Общий взгляд на этапы и задачи

Современный жизненный цикл КИС с ИИ имеет специфические черты, связанные с работой с данными и моделями и тесную связь с задачами инфраструктуры и ИБ



История организационного и технологического развития

От импортозамещения Highload-стека к доверенной ИТ-инфраструктуре на ПАК

2014-2015 | 2016 | 2017 | 2019 | 2021 | 2023 | 2025

- **Замещение VMware и Citrix**
  Virtuozzo + VDI
  → **Динамическая инфраструктура**
  Машины виртуализации Скала^р МВ
- **Замещение Oracle Exadata**
  Postgres Professional
  → **Высокопроизводительные СУБД**
  Машины баз данных Скала^р МБД.П
- **Замещение Teradata**
  Arenadata
  → **Управление большими данными**
  Машины больших данных Скала^р МБД.8
- **Замещение Scality, EMC/Netapp**
  Virtuozzo SDS + S3
  → **Масштабные S3-хранилища**
  Машины хранения данных Скала^р МХД.О
- **Модульная платформа Скала^р**
  Приоритетное направления развития
  → **Развитие Модульной Платформы**
- **Инфраструктура для ИИ**
  → Машины искусственного интелекта Скала^р

22 мая 2015
первое публичное представление платформы Скала^р

Спасибо за внимание!

www.skala-r.ru

Документ актуализирован 02.10.2025