



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



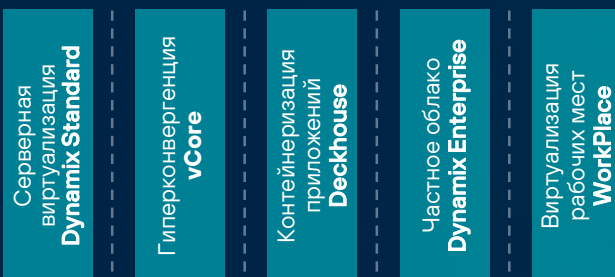
Мультитенантность



Автоматизация

Модульная платформа

Динамическая инфраструктура



Динамическая инфраструктура

Инфраструктура управления данными



Транзакционная обработка

Большие данные

Специализир. решения

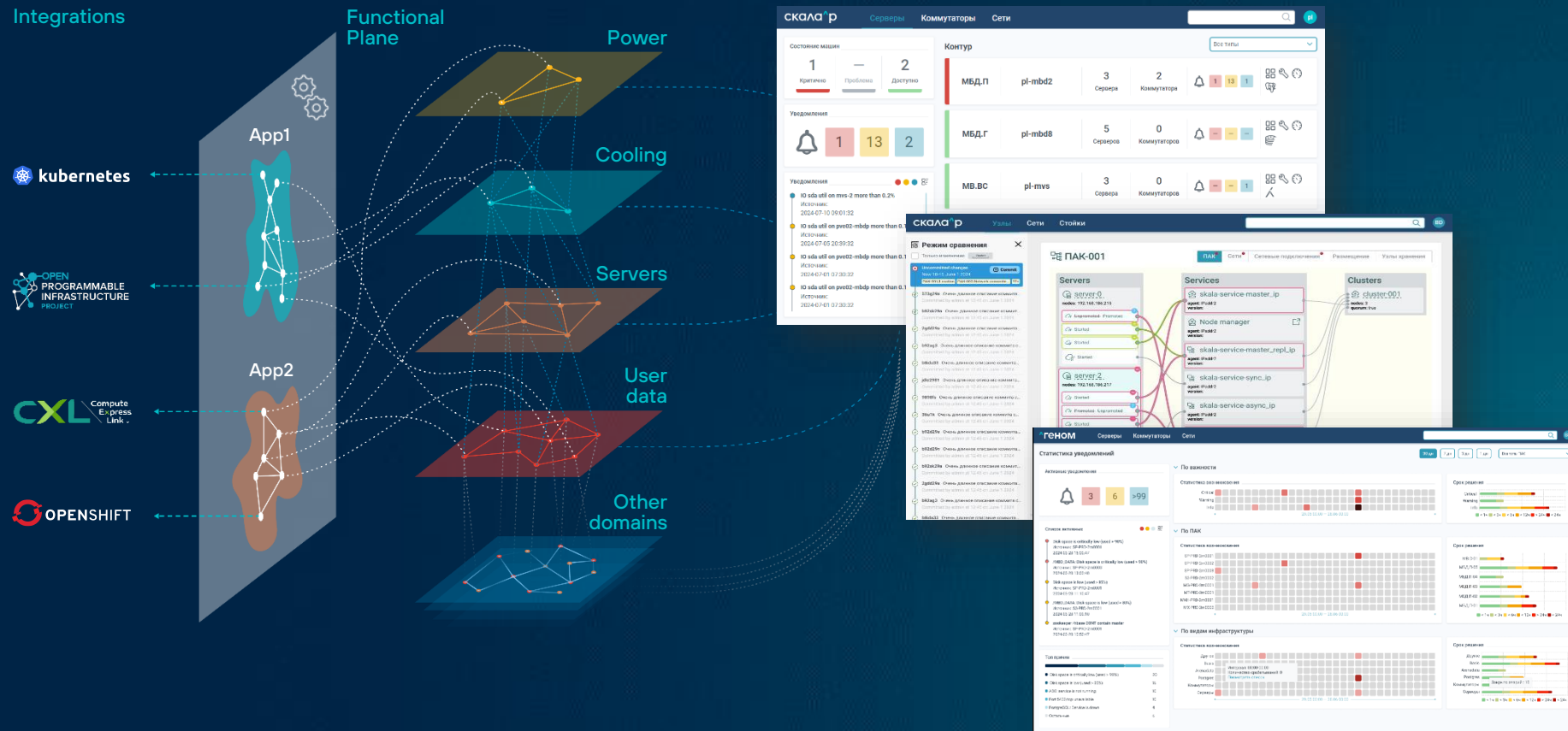
Отраслевые решения

Программная платформа Скала^р



Объединения различных доменов управления в единую функциональную графовую CMBD

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



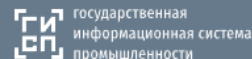
ПАК Скала[^]р в Реестрах РФ



Машины

Модули

Компоненты



Все сервисы ГИСП

Реестр промышленной продукции, произведенной на территории Российской Федерации

Машины

Модули

Программное обеспечение



РЕЕСТР
ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Российский

Евразийский

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



Хранение данных



Виртуализация



Вычислительная инфраструктура



СУБД



Операционные системы



Российский ИТ-рынок

Сетевая инфраструктура



Хранение данных



Виртуализация



Вычислительная инфраструктура



СУБД



Операционные системы



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

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

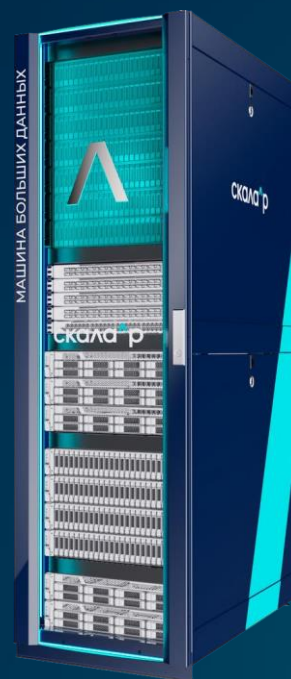
Экономия до 90%
на проектировании и внедрении

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

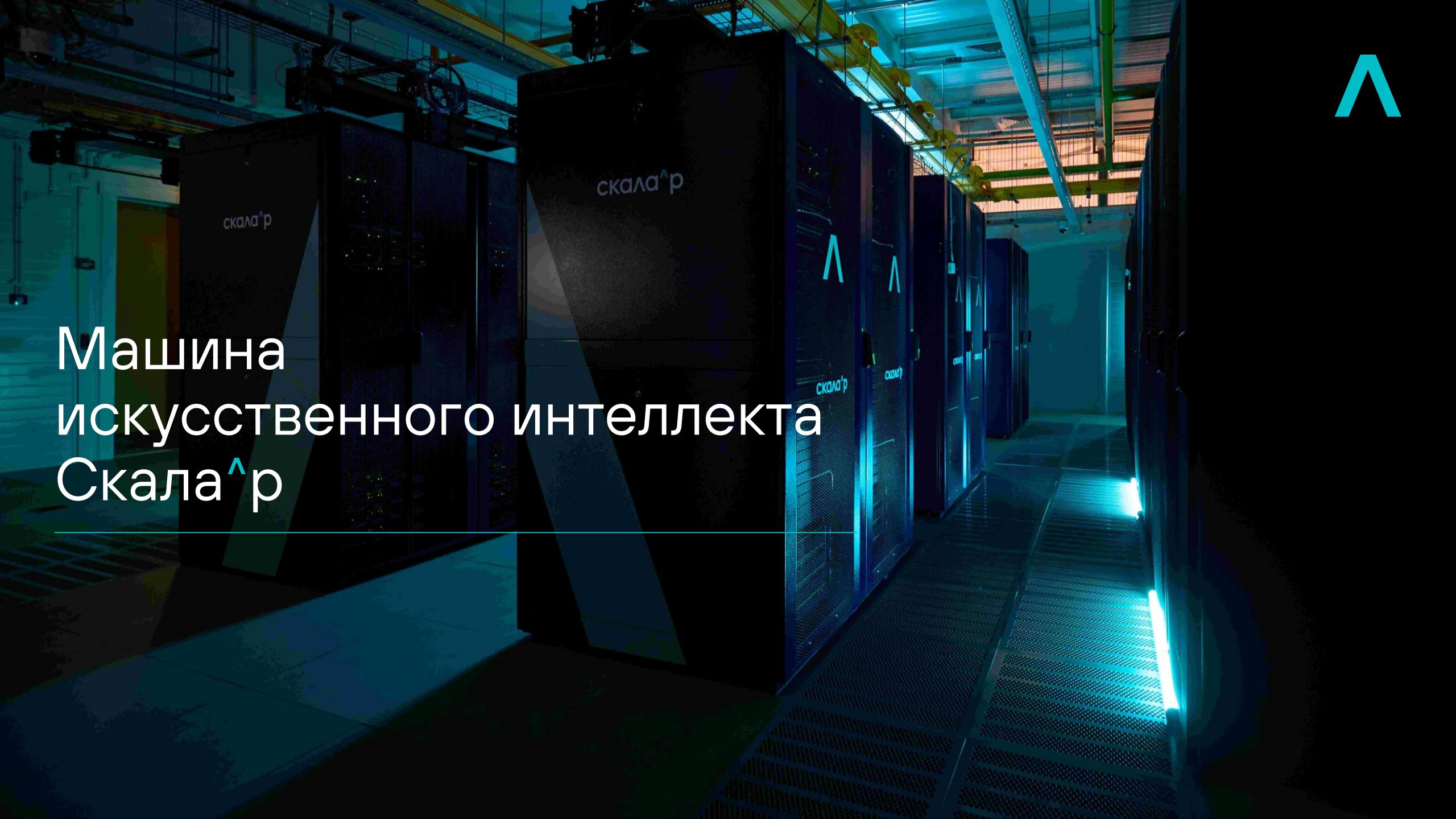
Удобство закупочных процедур для ПАК и Модулей —
это номенклатурные позиции Реестра РЭП
Минпромторга РФ

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 возвращения 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 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
- ИТ-специалисты -> MLDevOps -> ML инженеры -> DataScience

- Экономия на дальнейшей масштабируемости за счёт применения HPC технологий и устранения 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 задаче

- 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



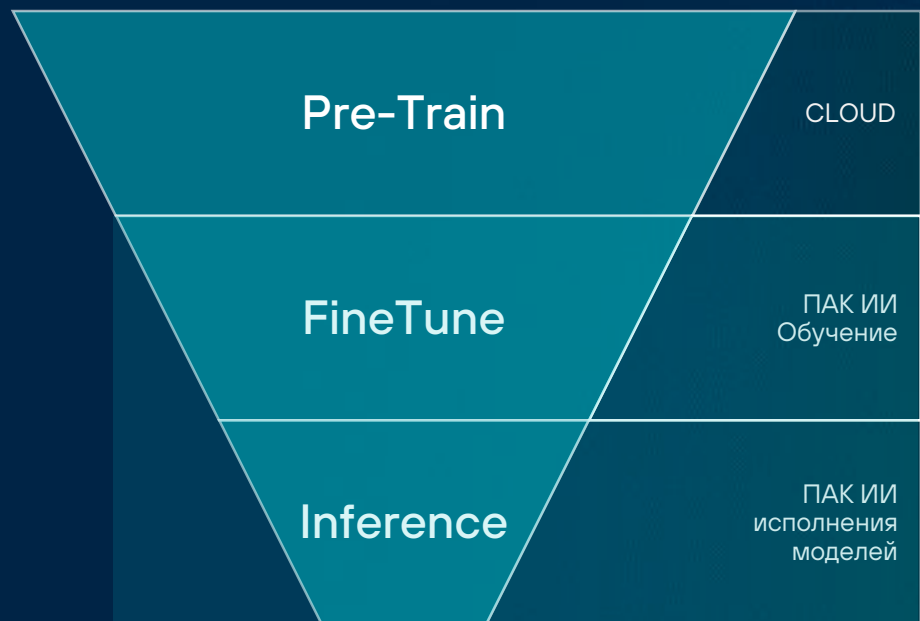
YandexGPT 5



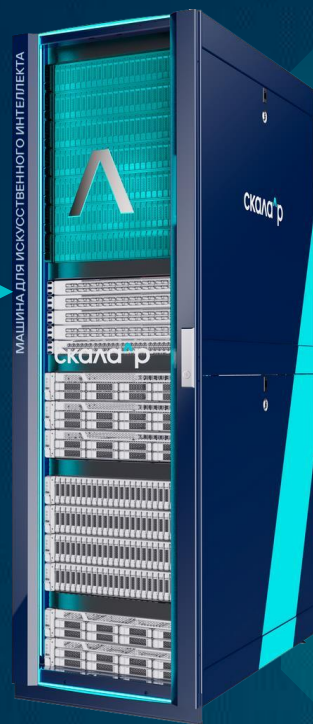
Смарт
Платформа



Другие
ИИ решения



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

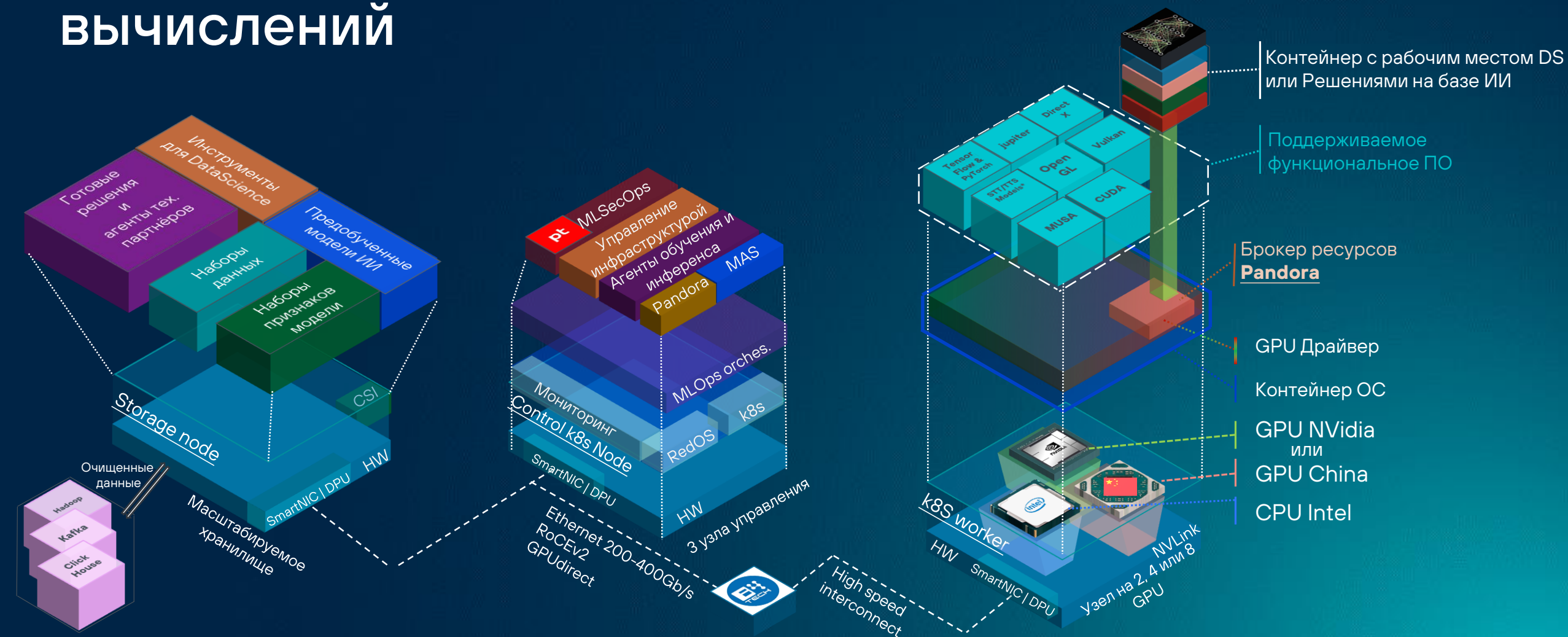


Аналог

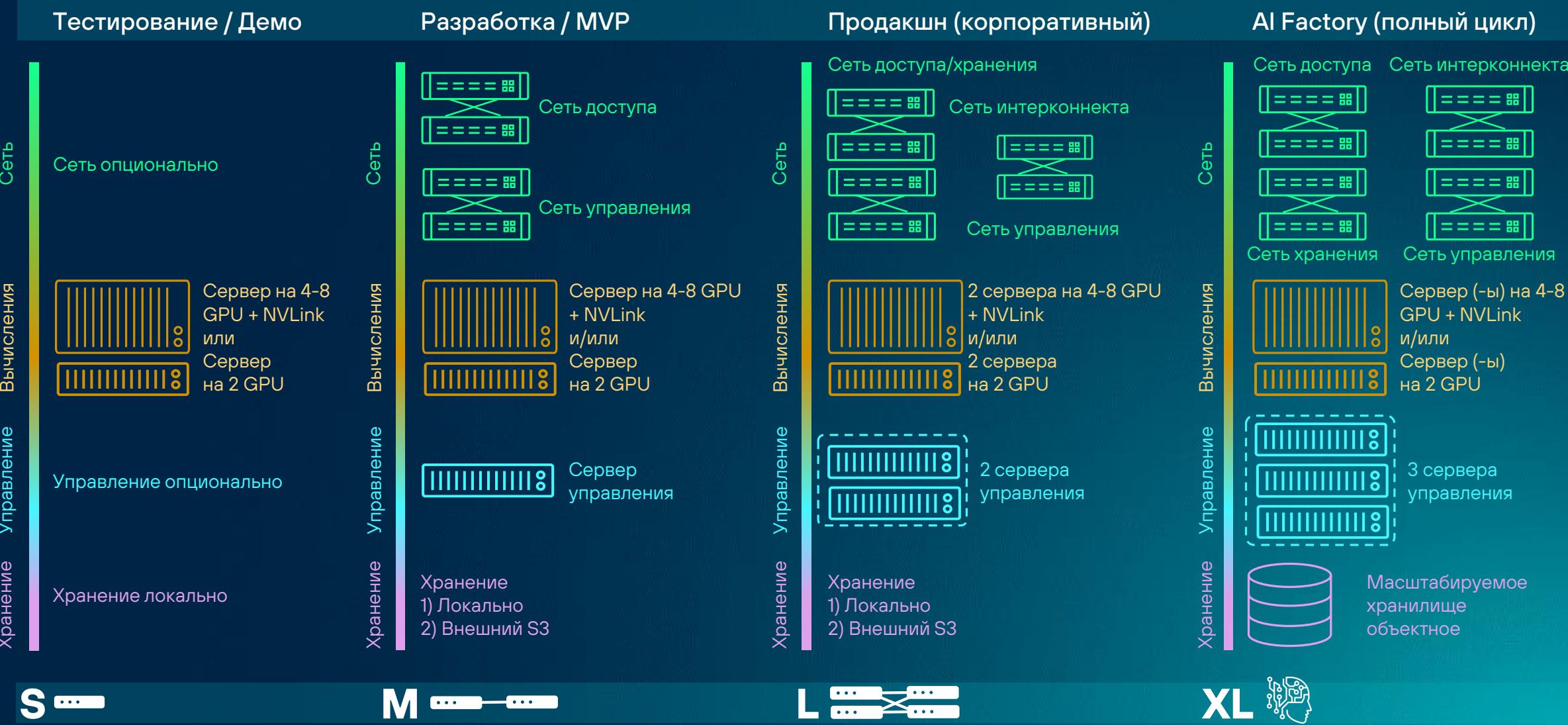
Huawei Atlas 900 Pod / NVIDIA DGX SuperPOD



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



Размеры Машины Скала^р МИИ



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 (Инструмент)



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 (Инструмент)



RAG (Инструмент)



Cotype (LLM)



DeepSeek (LLM)



ValueAI (Инструмент)



Llama (LLM)



GigaChat (LLM)



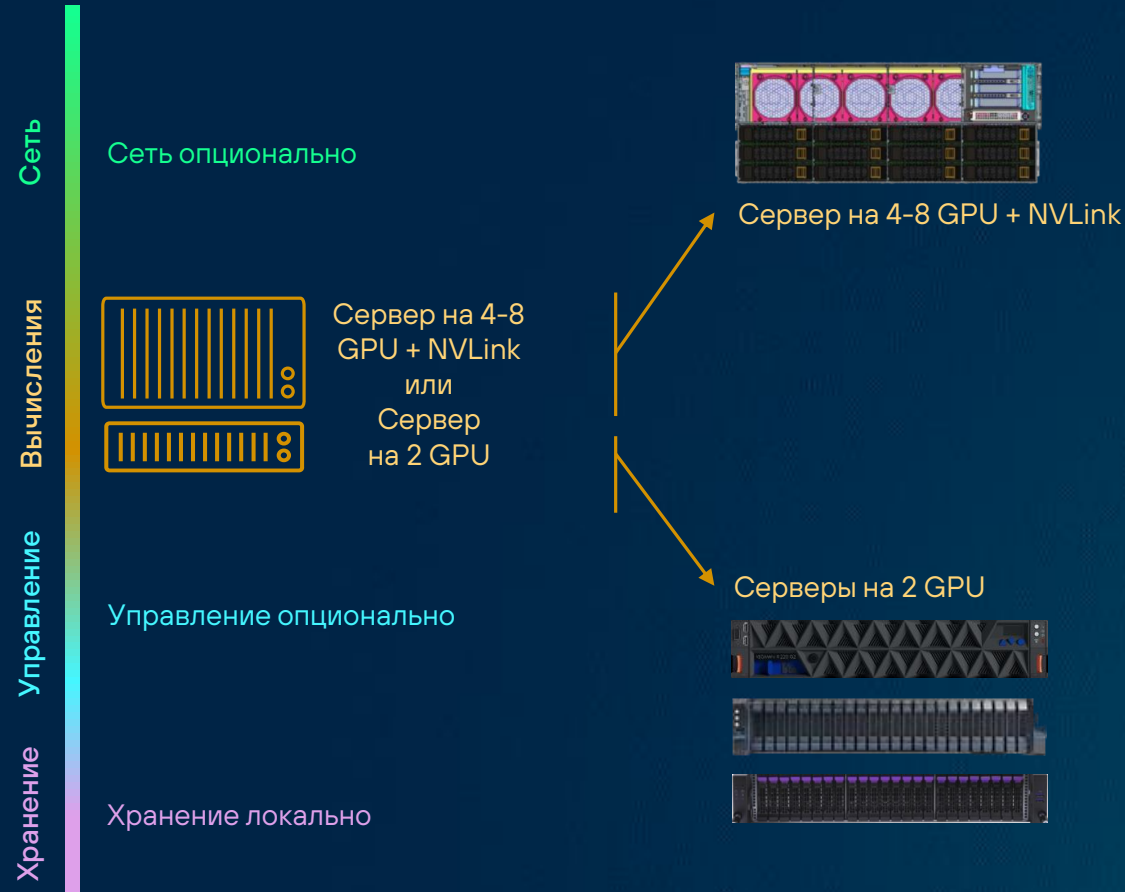
Другие ИИ решения



- 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

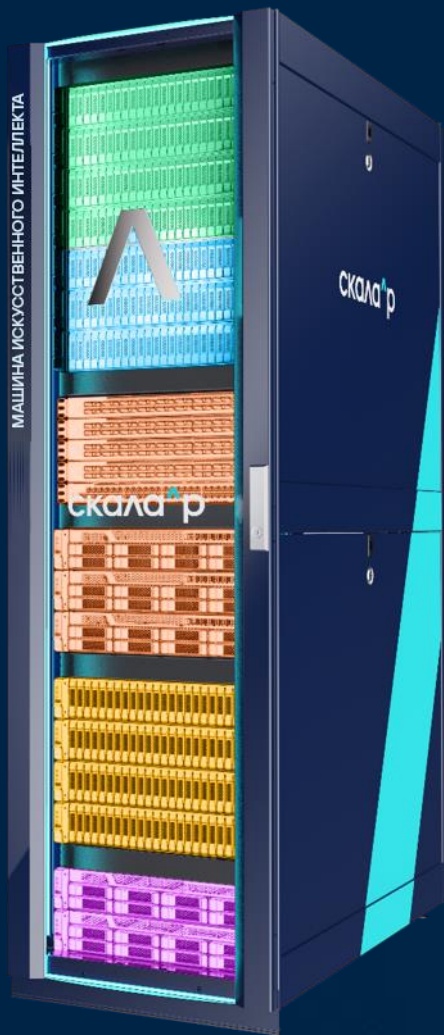


Юниты	4 RU
Процессоры	2x Intel Xeon 4/5 Gen
RAM	DDR5
PCIe слоты	Передняя панель: поддерживает максимум один слот PCIe 5.0 Задняя панель: поддерживает максимум 10 слотов PCIe 5.0 и 8 видеокарт двойной ширины
Электропитание	~3,6КВт

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).

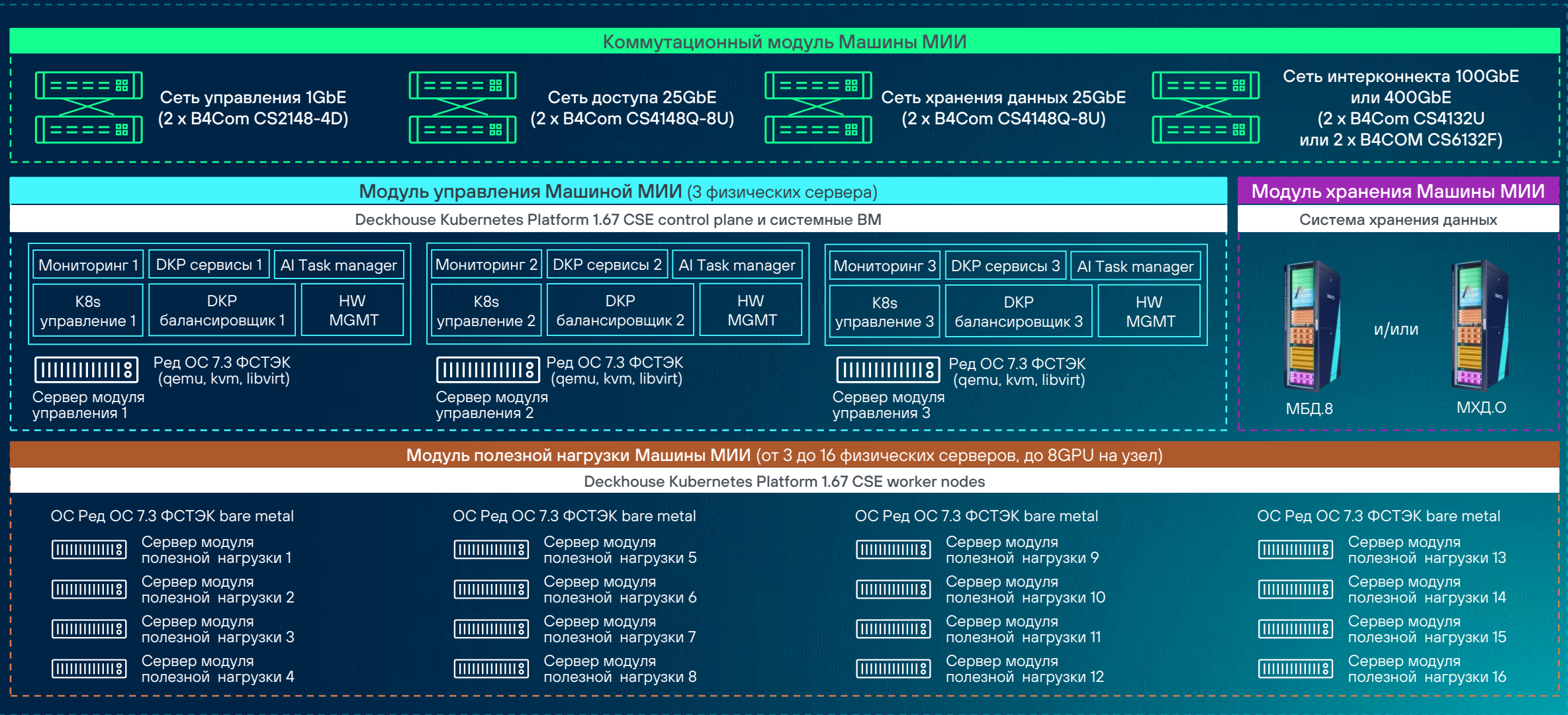
Модуль управления Машиной МИИ

- 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.

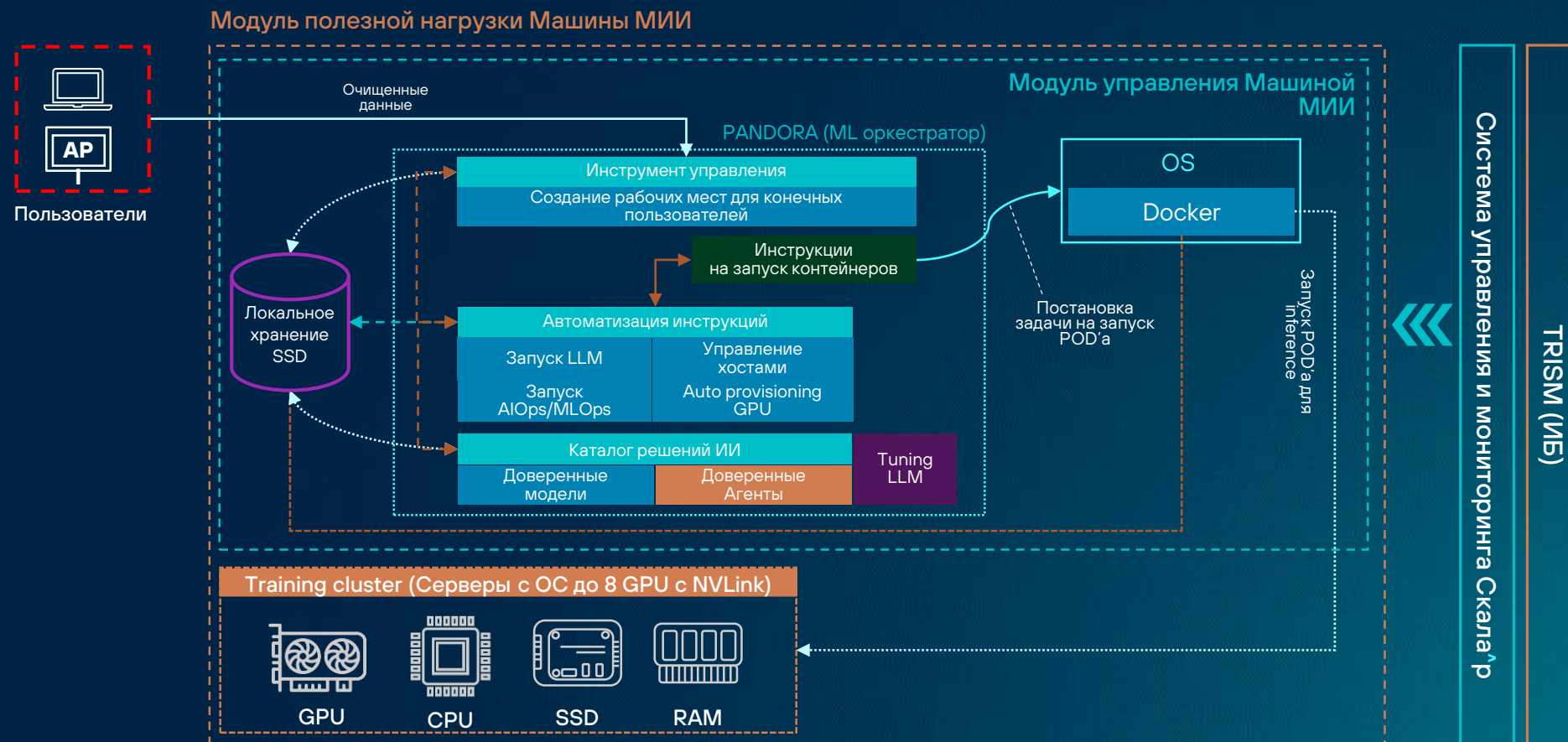
Модуль хранения Машины МИИ

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

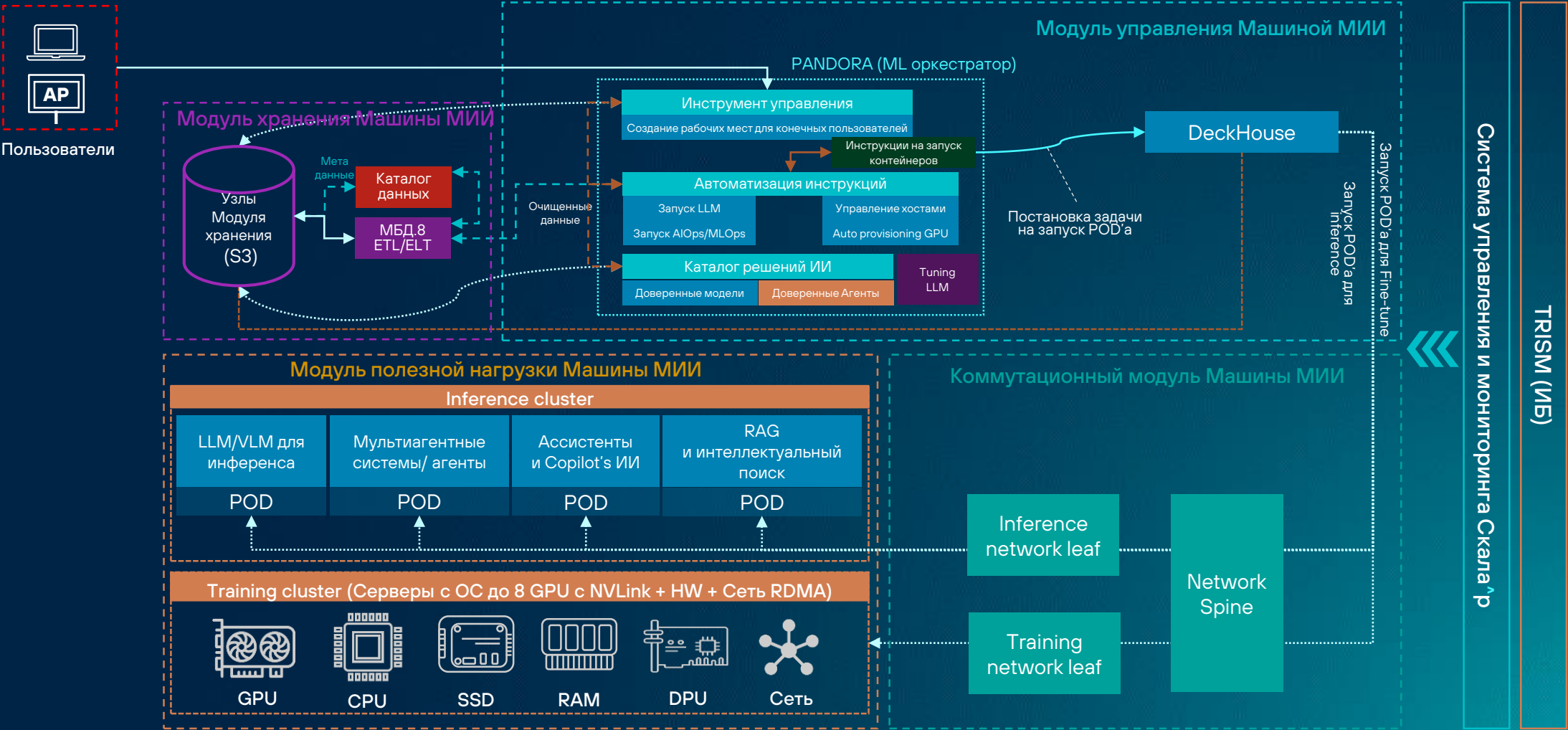
Машина Скала[^]р МИИ XL — компоненты



Машина Скала[^]р МИИ «S» — логическая схема



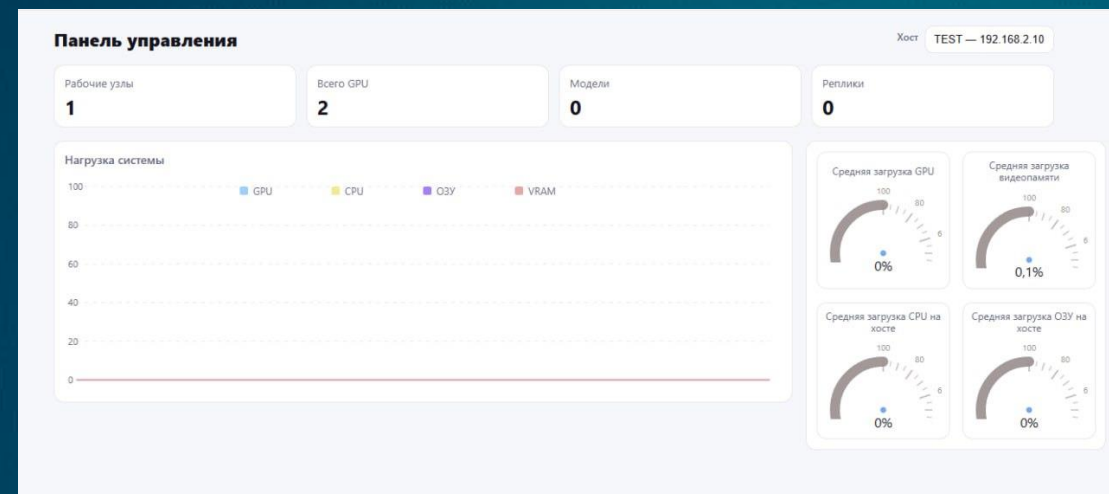
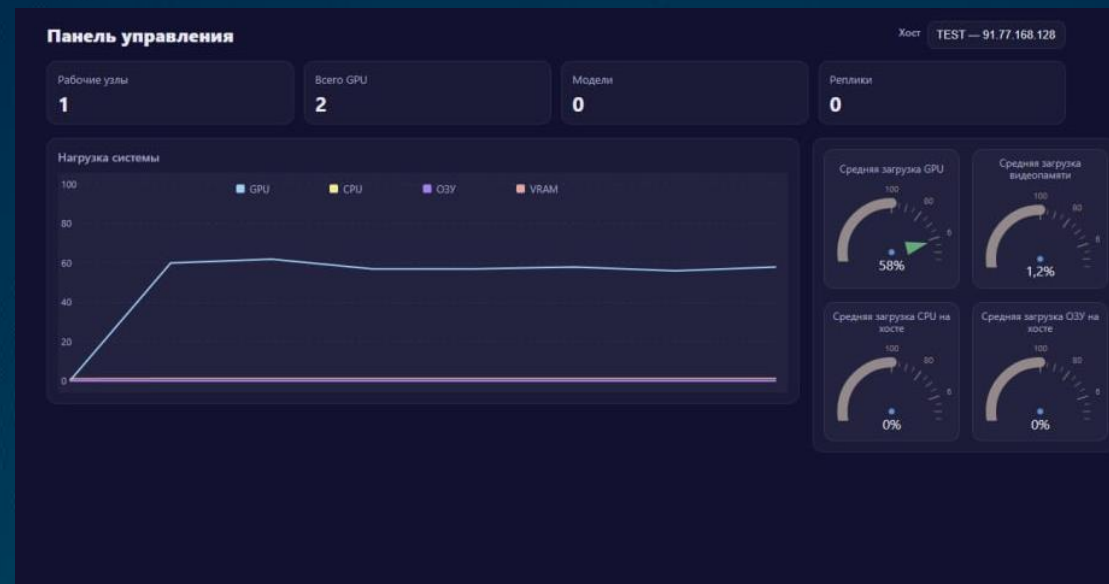
Машина Скала[^]р МИИ «XL» — логическая схема



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





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

Pandora

Панель управления

Создание контейнеров

Каталог

Deployment

Workers

GPUs

Users

ИИ

Настройки

Тема

Выйти

Вошли как admin (admin)

Создать контейнер

Создать контейнер

Хост

TEST — 91.77.168.128:2226

CPU (ядра)

4

Образ

redos8-mtt-kde:latest (9.14GB)

Протоколы

☒ noVNC

☐ RDP

☐ HTTP

Порты: noVNC 6080+, RDP 3390+, HTTP 8080+ (свободные на хосте)

Запустить контейнер

GPU

— без GPU —

RAM (ГБ)

8

Пользователь

— выберите —

Рабочий стол

Обновить

redos8-mtt-latest-moore_threads-084639

redos8-mtt:latest

@usertest, @port:6082

running

Подключиться

Перезапуск

Остановить

Удалить

nvidia-redos7-latest-nvidia-083702

nvidia-redos7:latest

@usertest, @port:6081

running

Подключиться

Перезапуск

Остановить

Удалить

nvidia-redos7-latest-nvidia-083428

nvidia-redos7:latest

@useralik, @port:6080

running

Подключиться

Перезапуск

Остановить

Удалить



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

Pandora

Панель управления

Создание контейнеров

Каталог

Deployment

Workers

GPUs

Users

ИИ

Чат

Генерация изображений

MLOps

Настройки

Тема Выйти

Вошли как admin (admin)

Каталог

Категория LLM Хост TEST — 192.168.2

Фильтр по имени Все

meta-llama-llama-3....
context-labs
12837232
4

2025-02-22 llama3.2 LLM

38

t5gemma-b-b-pre...
google
11666949
7

2025-07-09 gemma LLM

gpt2
openai-community
11136700
2932

2024-02-19 mit LLM

opt-125m
facebook
10503740
214

2023-09-15 other LLM

Qwen2.5-7B-Ins...
Qwen
10060681
785

2025-01-12 apache-2.0 LLM

78

gpt-oss-20b
openai
8551993
3479

2025-08-26 apache-2.0 LLM

208

Llama-3.1-8B-In...
meta-llama
7297191
4597

2024-09-25 llama3.1 LLM

88

Llama-3.2-1B-In...
meta-llama
7261922
1061

2024-10-24 llama3.2 LLM

18

Qwen2.5-3B-Inst...
Qwen
5158945
304

2024-09-25 other LLM

38

Qwen3-0.6B
Qwen
4789337
613

2025-07-26 apache-2.0 LLM

0.6B

Qwen2.5-1.5B-In...
Qwen
4723914
509

2024-09-25 apache-2.0 LLM

1.5B

dolphin-2.9.1-yl-1...
dolphin
4701802
39

2025-09-08 apache-2.0 LLM

348

tiny-Qwen2ForCaus...
tri-internal-testing
4375178
1

2025-08-02 LLM

Qwen2.5-0.5B-Ins...
Gensyn
4115798
21

2025-03-31 apache-2.0 LLM

0.5B

distilgpt2
distilbert
3843277
570

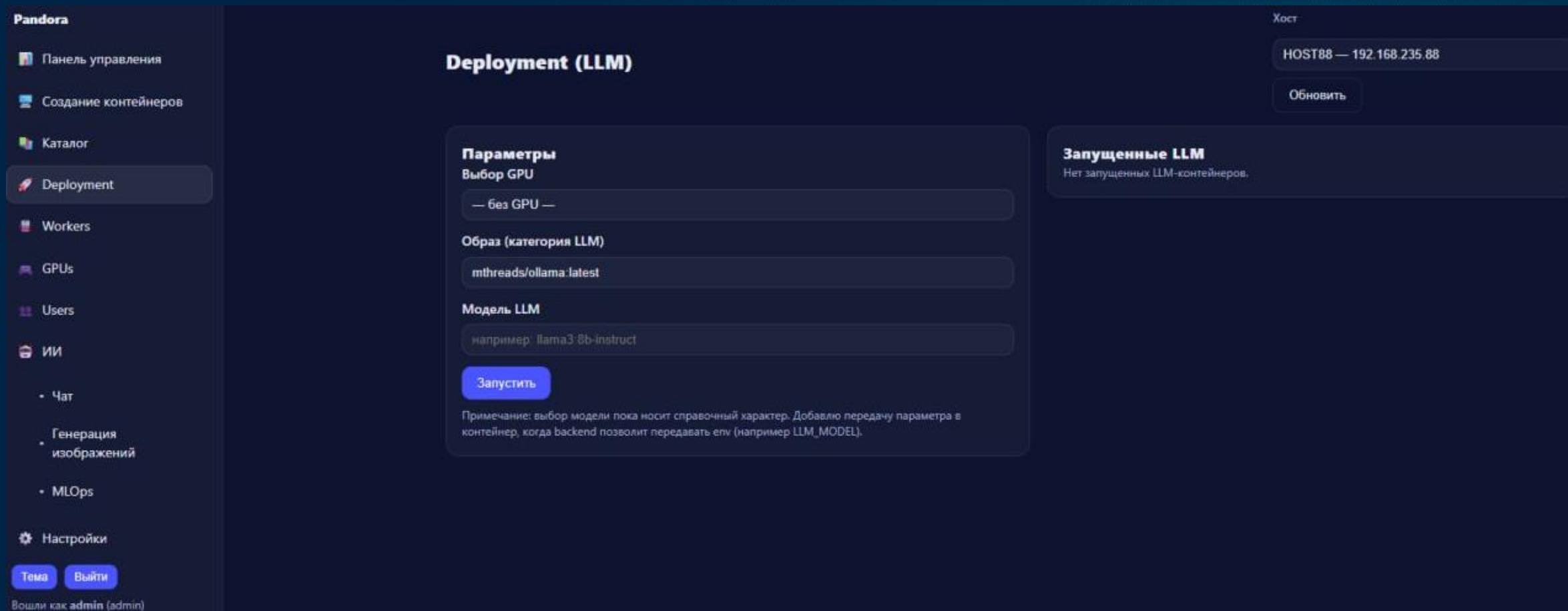
2024-02-19 apache-2.0 LLM

Llama-3.2-1B
meta-llama
3333926
2077

2024-10-24 llama3.2 LLM

1B

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



The screenshot displays the Pandora web interface for LLM deployment. On the left is a dark sidebar with navigation links: 'Панель управления', 'Создание контейнеров', 'Каталог', 'Deployment' (highlighted), 'Workers', 'GPUs', 'Users', 'ИИ' (with sub-items 'Чат', 'Генерация изображений', 'MLOps'), and 'Настройки'. At the bottom of the sidebar are 'Тема' and 'Выйти' buttons, and a login status 'Вошли как admin (admin)'. The main content area is titled 'Deployment (LLM)'. It features a 'Хост' section with 'HOST88 — 192.168.235.88' and an 'Обновить' button. Below this is the 'Запущенные LLM' section, which currently shows 'Нет запущенных LLM-контейнеров.' The central 'Параметры' section includes a 'Выбор GPU' dropdown (set to '— без GPU —'), an 'Образ (категория LLM)' dropdown (set to 'mthreads/ollama:latest'), and an 'Модель LLM' text input (containing 'например: llama3-8b-instruct'). A blue 'Запустить' button is positioned below these inputs. A note at the bottom of the parameters section states: 'Примечание: выбор модели пока носит справочный характер. Добавлю передачу параметра в контейнер, когда backend позволит передавать env (например LLM_MODEL)'.

PANDORA



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

The screenshot displays the Pandora web interface. On the left is a dark sidebar with a menu. The main content area is titled 'GPUs' and shows a list of available GPU resources. The sidebar menu includes: Pandora, Панель управления, Создание контейнеров, Каталог, Deployment, Workers, GPUs (highlighted), Users, ИИ, Чат, Генерация изображений, MLOps, and Настройки. At the bottom of the sidebar are buttons for 'Тема' and 'Выйти', and a login status 'Вошли как admin (admin)'. The main area has a 'Хост' label with the value 'TEST — 91.77.168.128'. Under the 'GPUs' heading, there are two sections: 'NVIDIA' with one item 'NVIDIA RTX A4000 • GPU-c12b47c1-cc9f-6445-3513-ca6fe996d94c', and 'MOORE_THREADS' with one item 'MTT S80 • c2638ab9-e4ea-89d1-3867-f9e47643f448'.

Pandora

- Панель управления
- Создание контейнеров
- Каталог
- Deployment
- Workers
- GPUs**
- Users
- ИИ
 - Чат
 - Генерация изображений
 - MLOps
- Настройки

Тема Выйти

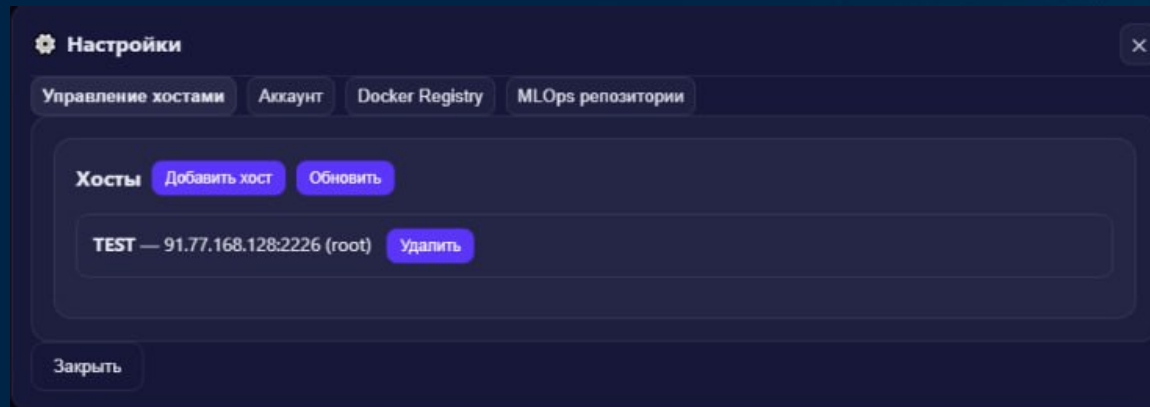
Вошли как **admin** (admin)

GPUs

Хост: TEST — 91.77.168.128

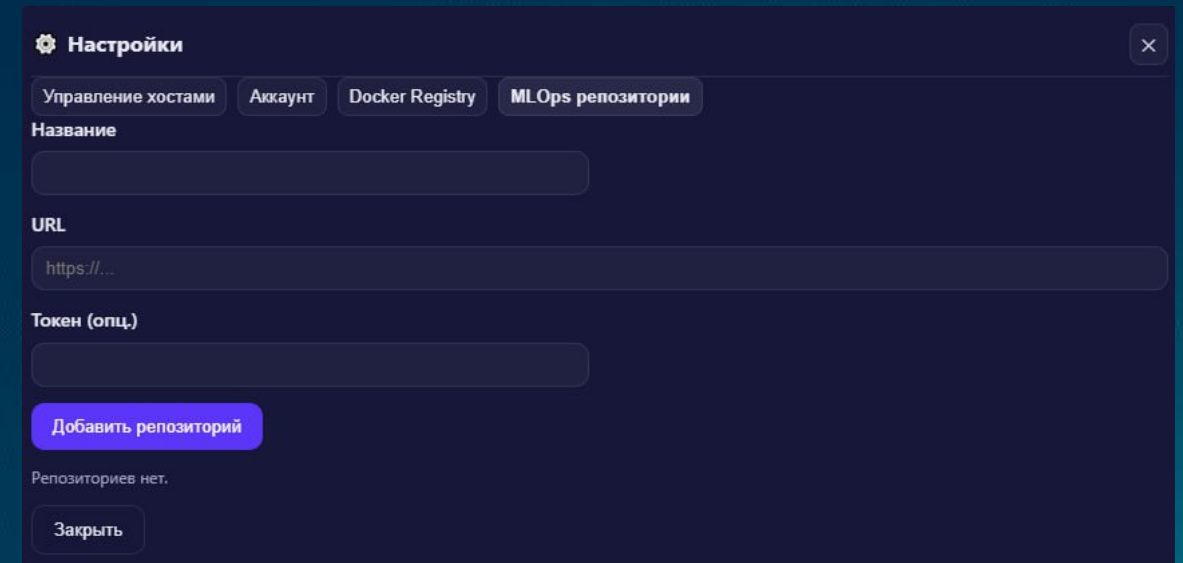
- NVIDIA**
 - NVIDIA RTX A4000 • GPU-c12b47c1-cc9f-6445-3513-ca6fe996d94c
- MOORE_THREADS**
 - MTT S80 • c2638ab9-e4ea-89d1-3867-f9e47643f448

Модальное окно управления хостом (-ами)



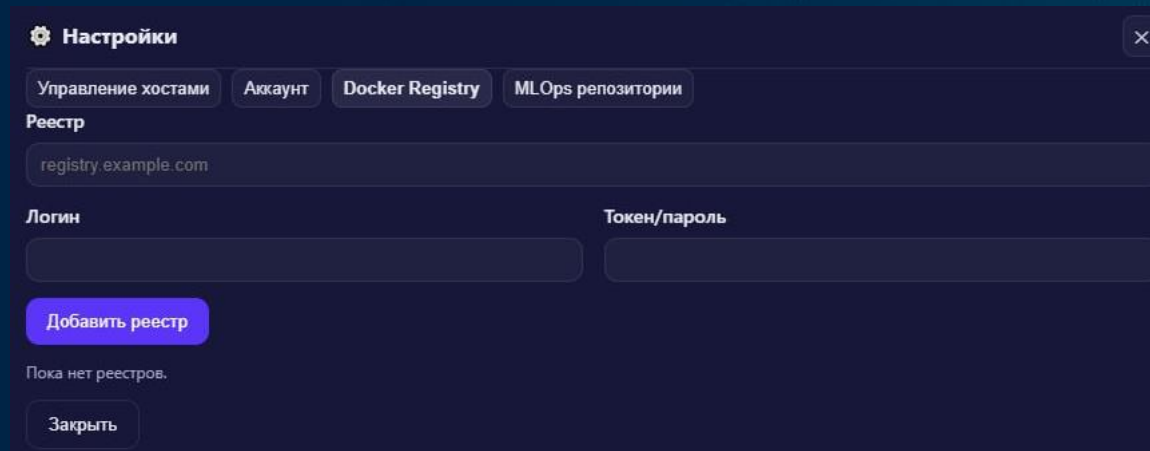
The screenshot shows a modal window titled "Настройки" (Settings) with a close button (X) in the top right corner. Below the title bar, there are four tabs: "Управление хостами" (Host Management), "Аккаунт" (Account), "Docker Registry", and "MLOps репозитории" (MLOps Repositories). The "Управление хостами" tab is selected. Inside this tab, there is a section labeled "Хосты" (Hosts) with two buttons: "Добавить хост" (Add host) and "Обновить" (Refresh). Below this, there is a list of hosts. One host is visible: "TEST — 91.77.168.128:2226 (root)" with a "Удалить" (Delete) button next to it. At the bottom left of the modal, there is a "Закрыть" (Close) button.

Модальное окно добавления репозитория MLOps для LLM



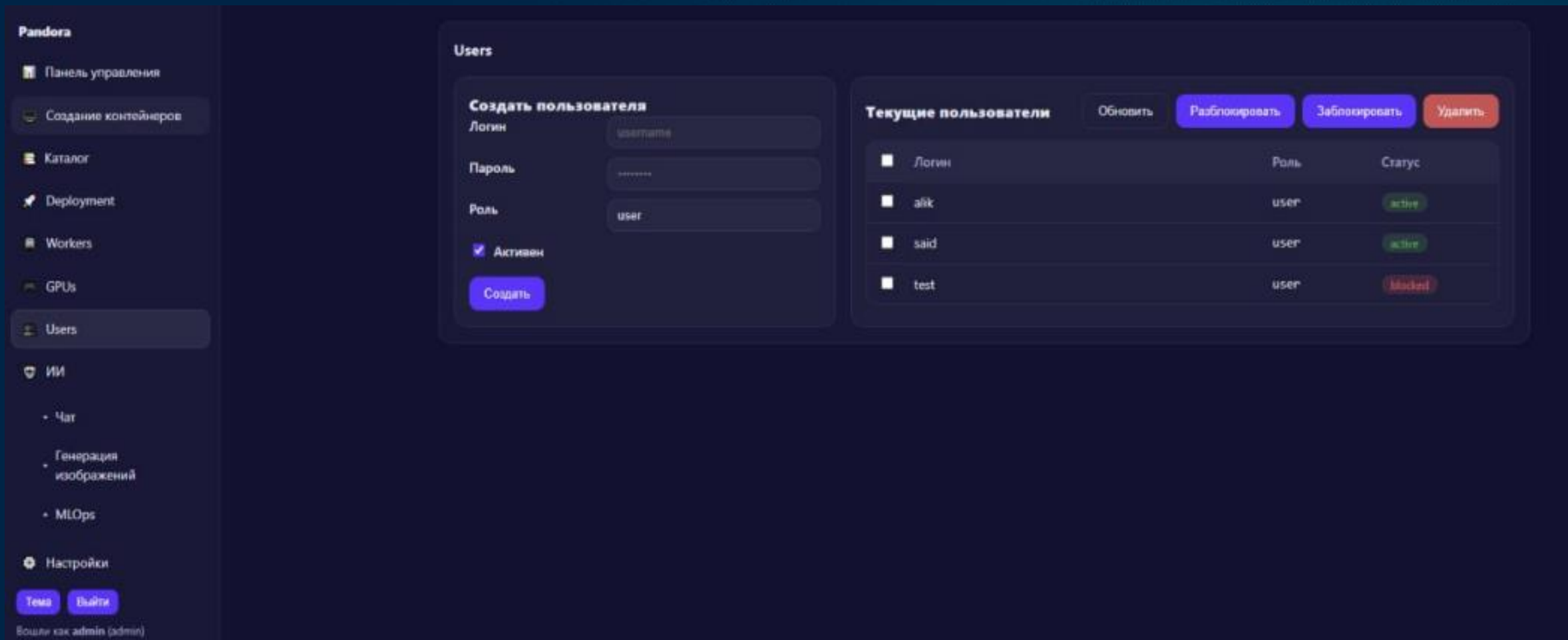
The screenshot shows a modal window titled "Настройки" (Settings) with a close button (X) in the top right corner. Below the title bar, there are four tabs: "Управление хостами" (Host Management), "Аккаунт" (Account), "Docker Registry", and "MLOps репозитории" (MLOps Repositories). The "MLOps репозитории" tab is selected. Inside this tab, there are three input fields: "Название" (Name), "URL", and "Токен (опц.)" (Token (optional)). Below these fields, there is a "Добавить репозиторий" (Add repository) button. At the bottom of the modal, there is a "Закрыть" (Close) button and a message "Репозитория нет." (Repository not found).

Модальное окно добавления репозитория docker



The screenshot shows a modal window titled "Настройки" (Settings) with a close button (X) in the top right corner. Below the title bar, there are four tabs: "Управление хостами" (Host Management), "Аккаунт" (Account), "Docker Registry", and "MLOps репозитории" (MLOps Repositories). The "Docker Registry" tab is selected. Inside this tab, there is a section labeled "Реестр" (Registry) with an input field for the registry name (e.g., "registry.example.com"). Below this, there are two input fields: "Логин" (Login) and "Токен/пароль" (Token/password). Below these fields, there is a "Добавить реестр" (Add registry) button. At the bottom of the modal, there is a "Закрыть" (Close) button and a message "Пока нет реестров." (No registries yet).

Интерфейс управления правами доступа



The screenshot displays the Pandora web interface for user management. On the left is a dark sidebar with a menu containing: "Панель управления", "Создание контейнеров", "Каталог", "Deployment", "Workers", "GPUs", "Users" (highlighted), "ИИ", "Чат", "Генерация изображений", "MLOps", and "Настройки". At the bottom of the sidebar are buttons for "Тема" and "Выйти", and a login status "Вошли как admin (admin)".

The main content area is titled "Users" and is divided into two panels:

- Создать пользователя (Create User):** A form with fields for "Логин" (username: "username"), "Пароль" (password: masked), and "Роль" (role: "user"). There is a checked checkbox for "Активен" (Active) and a blue "Создать" (Create) button.
- Текущие пользователи (Current Users):** A table listing existing users with action buttons above it.

Buttons above the table: "Обновить" (Update), "Разблокировать" (Unblock), "Заблокировать" (Block), and "Удалить" (Delete).

Логин	Роль	Статус
alik	user	active
said	user	active
test	user	blocked



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

Pandora

Панель управления

Создание контейнеров

Каталог

Deployment

Workers

GPUs

Users

ИИ

Чат

Генерация изображений

MLOps

Настройки

Тема

Выйти

Вошли как admin (admin)

Чат

Сравнение

</> Показать JSON

Системное сообщение

Краткая роль ассистента для сессии

Пользователь

кто ты

Ассистент

Я — Qwen, крупный язык модели, разработанный Alibaba Group. Я могу отвечать на вопросы, генерировать текст, помогать с задачами и поддерживать диалог. Однако я не могу выполнять задания, требующие взаимодействия с внешними ресурсами или доступ к реальному времени. Информация, которую я предоставляю, основана на моем обучении данных до 2023 года. Если у вас есть конкретные вопросы, я стараюсь помочь!

Пользователь

что ты можешь ?

Ассистент

Я — модель, которая может выполнять множество задач, включая: 1. **Генерацию текста** (писательство, написание сочинений, рассказов, статей). 2. **Ответы на вопросы** (по различным темам, от науки до культуры). 3. **Разбор текста** (анализ структуры, смысл, синтаксис). 4. **Помощь с задачами** (решение математических задач, написание кода, составление планов). 5. **Поддержку диалога** (обсуждение тем, интересующих пользователей). 6. **Создание презентаций** (представление информации в виде слайдов). 7. **Обучение и поддержка** (помощь в изучении языка, решении сложных задач). Могу ли я помочь вам с конкретной задачей? 😊

Ваш запрос

Напишите сообщение...

+ Сообщение

Send

Параметры

Хост

TEST — 192.168.2.10

Контейнер

gpustack/gpustack:latest-cuda12.8

GPUStack Base URL

http://192.168.2.10/v1

GPUStack API Key

gpustack_aebdbff5592c3f7_4665a226a2157f558af

Модель

Обновить

qwen3

Temperature

1

Max Tokens

0

Top P

1

Frequency Penalty

0

Presence Penalty

0

Seed

Stop Sequence

lnln

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

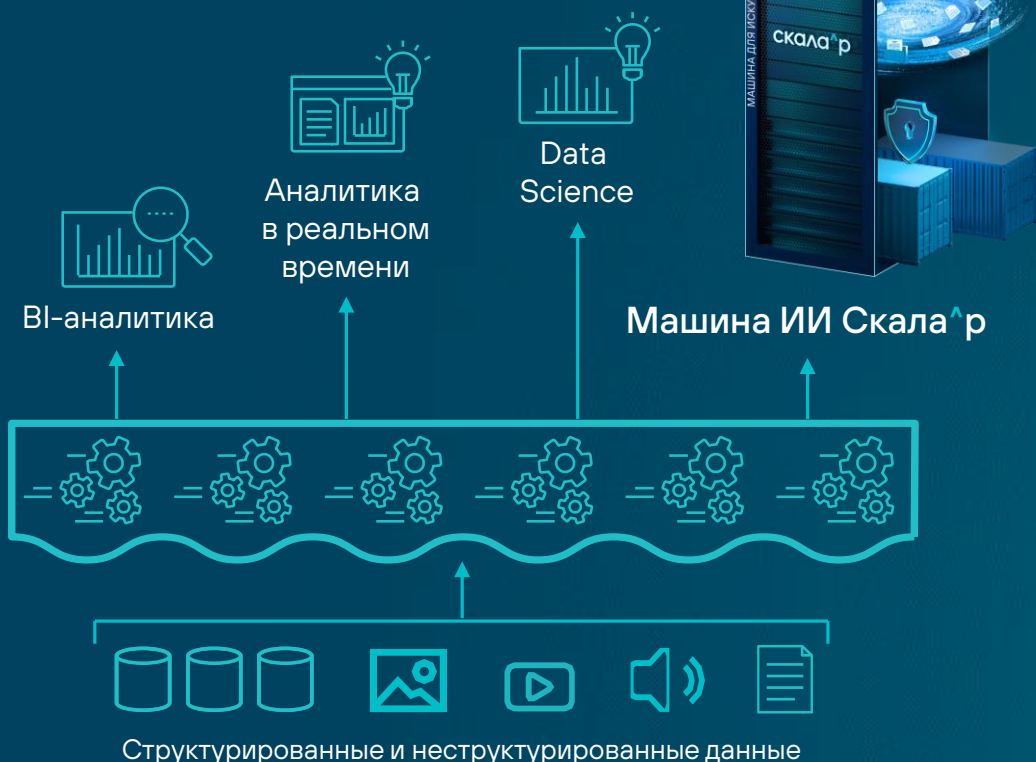
в 2–5 раз

в зависимости от количества пользователей и типа запросов

Экосистема данных Lakehouse + ML/AI



Платформа данных Data Lakehouse



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 моделей 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).

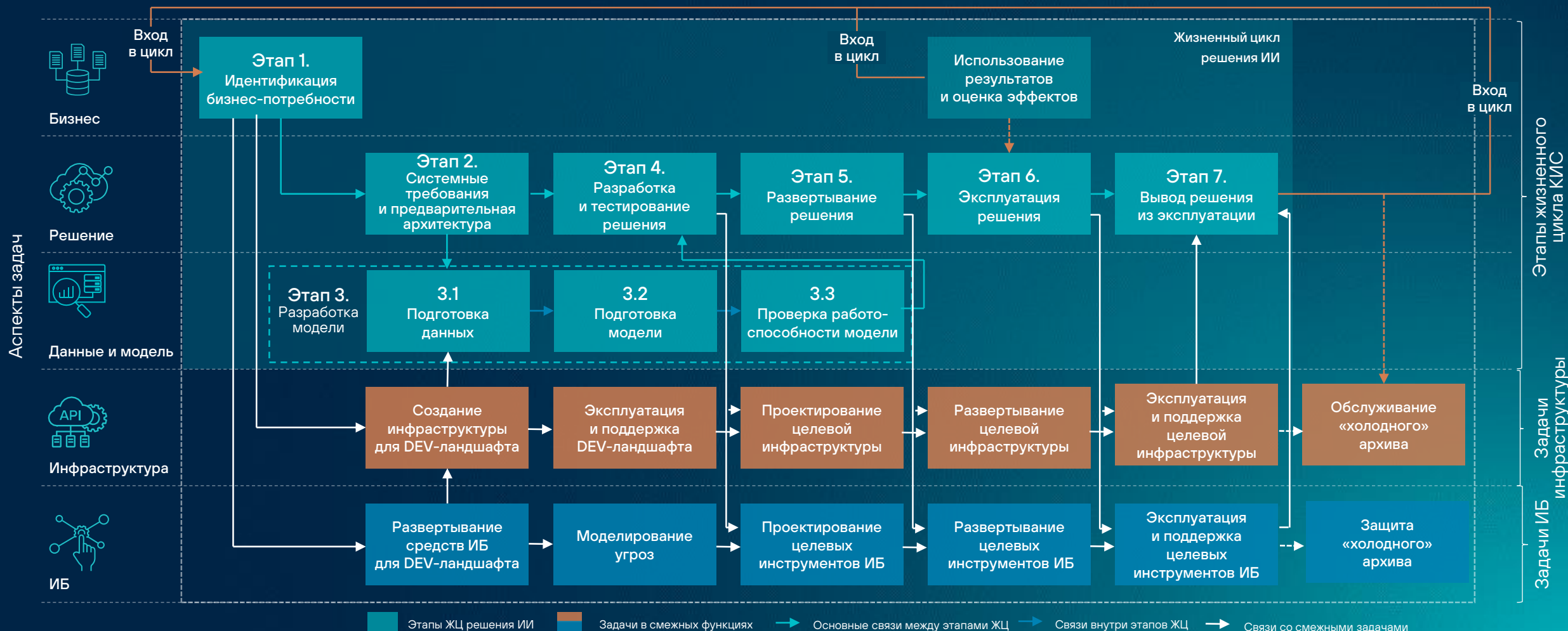
Построена на основе технологий NLP и применения адаптированных языковых моделей LLM.

* Типовые задачи для инфраструктуры Машины ИИ Скала[®]

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

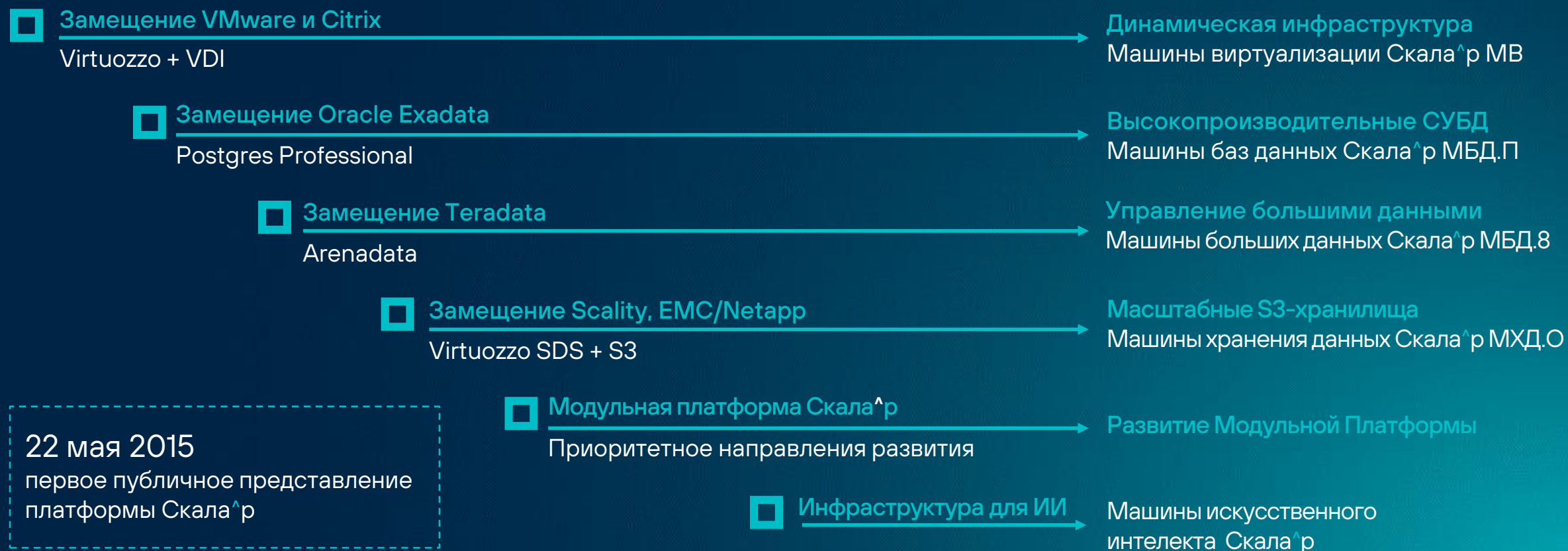


История организационного и технологического развития



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

2014-2015 2016 2017 2019 2021 2023 2025





Спасибо за внимание!



www.skala-r.ru