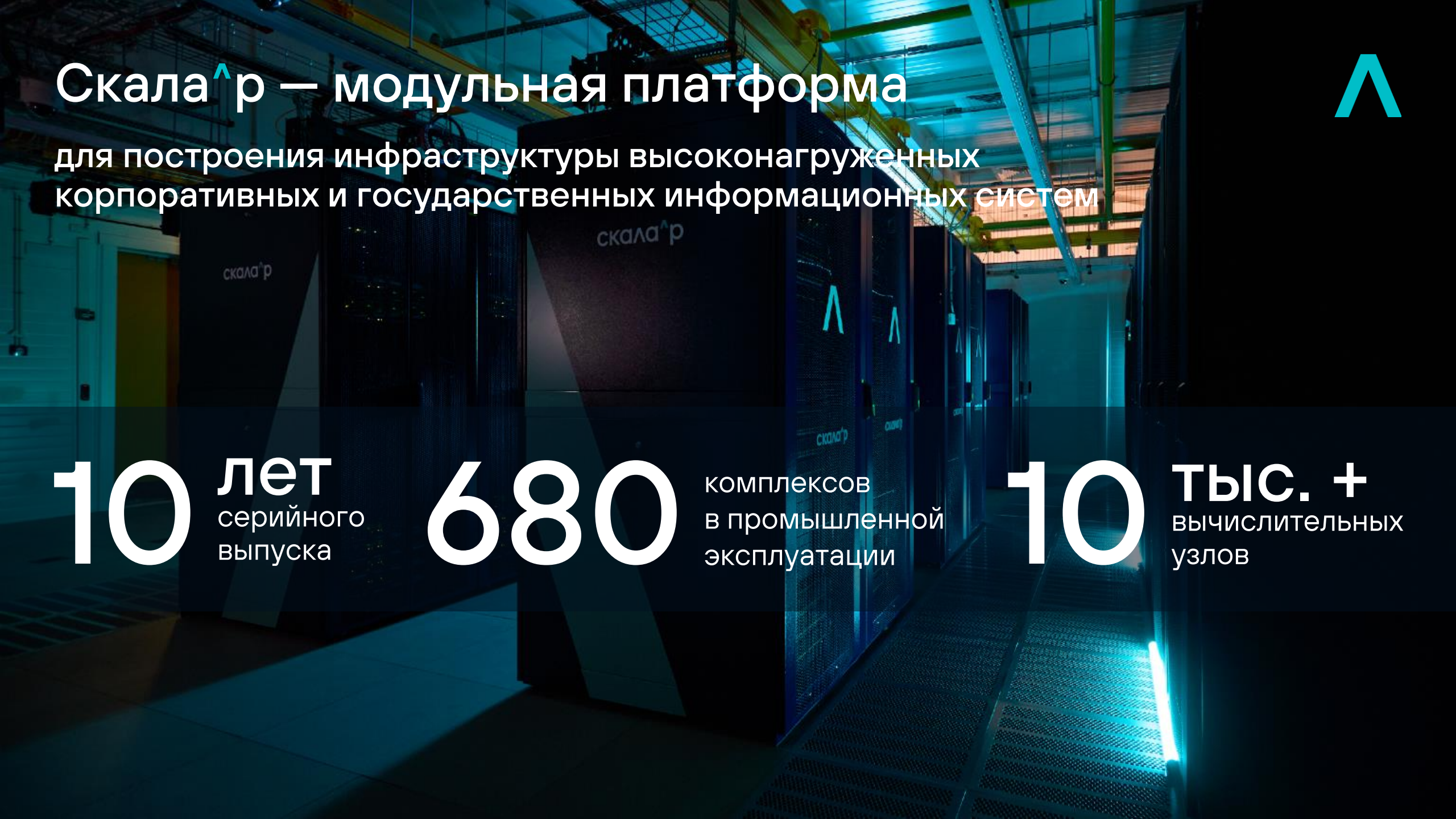
Машина искусственного интеллекта Скала^р



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

Инфраструктура ИИ

Машина искусственного интеллекта Скала^р

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

Управление данными

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

Машины больших данных Скала^р МБД

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

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
- решения на основе платформы S3 и российского ПО для комплексных задач резервного копирования и восстановления крупных массивов данных со встроенной иерархией хранения и обеспечением высокой доступности копий

Специализированные решения

Машина управления технологическими процессами Скала^р МСП.ТП (АСУ ТП)

Высоконадежная инфраструктура для различных АСУ ТП промышленных предприятий с высокими требованиям к отказоустойчивости и информационной безопасности. Соответствует требованиям ЗОКИИ, в том числе критериям к Доверенным ПАК

Машина автоматизированных банковских систем Скала^р МСП.БС

на платформе Машин Скала^р для задач класса АБС и процессинговых решений с поддержкой высокой транзакционной и аналитической нагрузки, сегментирования баз данных и обеспечения ИБ

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

Программная платформа Скала^р для управления ресурсами и эксплуатацией

Разделение ресурсов

Мультитенантность

Автоматизация

Модульная платформа

Динамическая инфраструктура

- Серверная виртуализация **Dynamix Standard**
- Гиперконвергенция **Dynamix Standard**
- 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

Интеллектуальное хранение

- Искусственный интеллект

ИИ

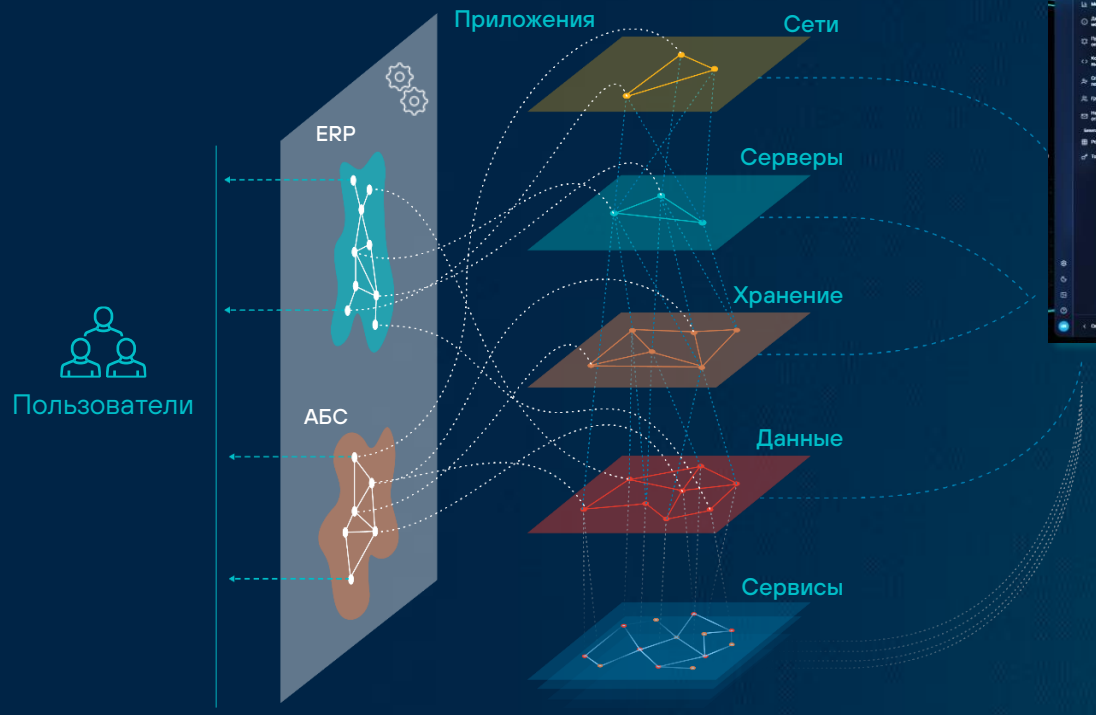
- Автоматизированные банковские системы
- АСУ ТП

Специализированные решения

Глубокая интеграция и встречная оптимизация компонентов по всему технологическому стеку под определенные нагрузки

Развитие: Программная платформа Скала^р

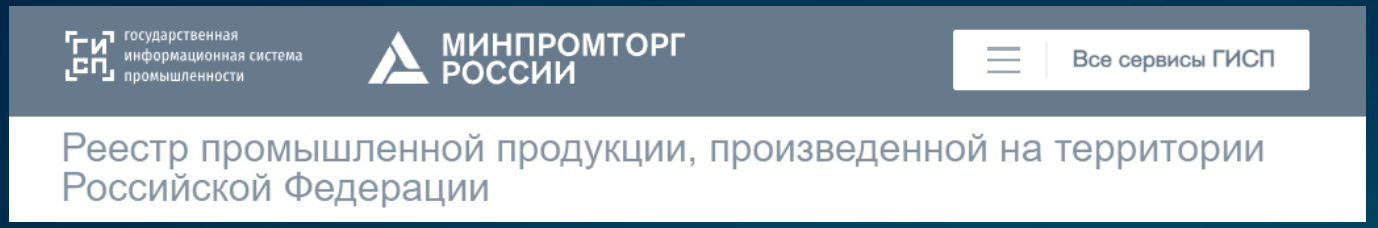
объединение различных доменов управления в единую объектно-сервисную графовую модель - комплексное решение для эксплуатации инфраструктуры уровня ЦОД



- Единая точка обзора состояния контура
- Обозримость и удобство управления ЦОД
- Цифровой двойник инфраструктуры
- Контроль изменений оборудования и сервисов
- Моделирование изменений в инфраструктуре
- Высокая степень автоматизации

ПАК Скала^р в Реестрах РФ

Машины (ПАК)
Модули (ПАК)
Компоненты



Машины (ПАК)
Модули (ПАК)
Программное обеспечение

ПАК Скала^р

Соответствуют критериям доверенного ПАК (ПП 1912)

- 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

Импортозамещение: варианты перехода

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 отсутствие необходимости ее масштабирования
- Сокращение сроков внедрения и снижение рисков

Почему ПАК Скала^р?

Высокая отказоустойчивость

За счет специализированной модульной и кластерной архитектуры решений

Высокая производительность

Встречная оптимизация и устранение узких мест по всему стеку применимых технологий

Единая техническая поддержка

Сопровождение оборудования и программного обеспечения всех компонентов Машин

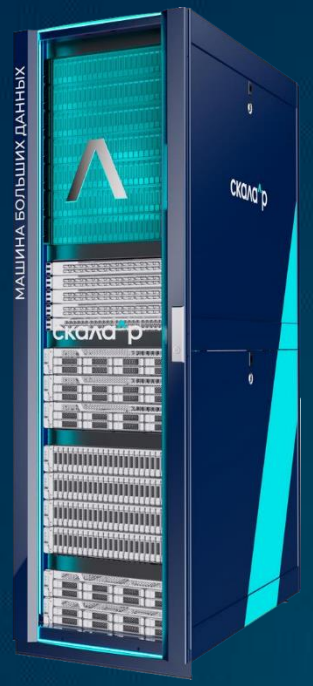
Экономия до 90% на **проектировании и внедрении**

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-аппаратные комплексы и Модули платформы — включены в Реестр российской промышленной и радиоэлектронной продукции, ПО Скала^р - в реестр Минцифры

Все ПАК Скала^р: встроенная безопасность

Внутренняя инфраструктура ПАК

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

Интеграция

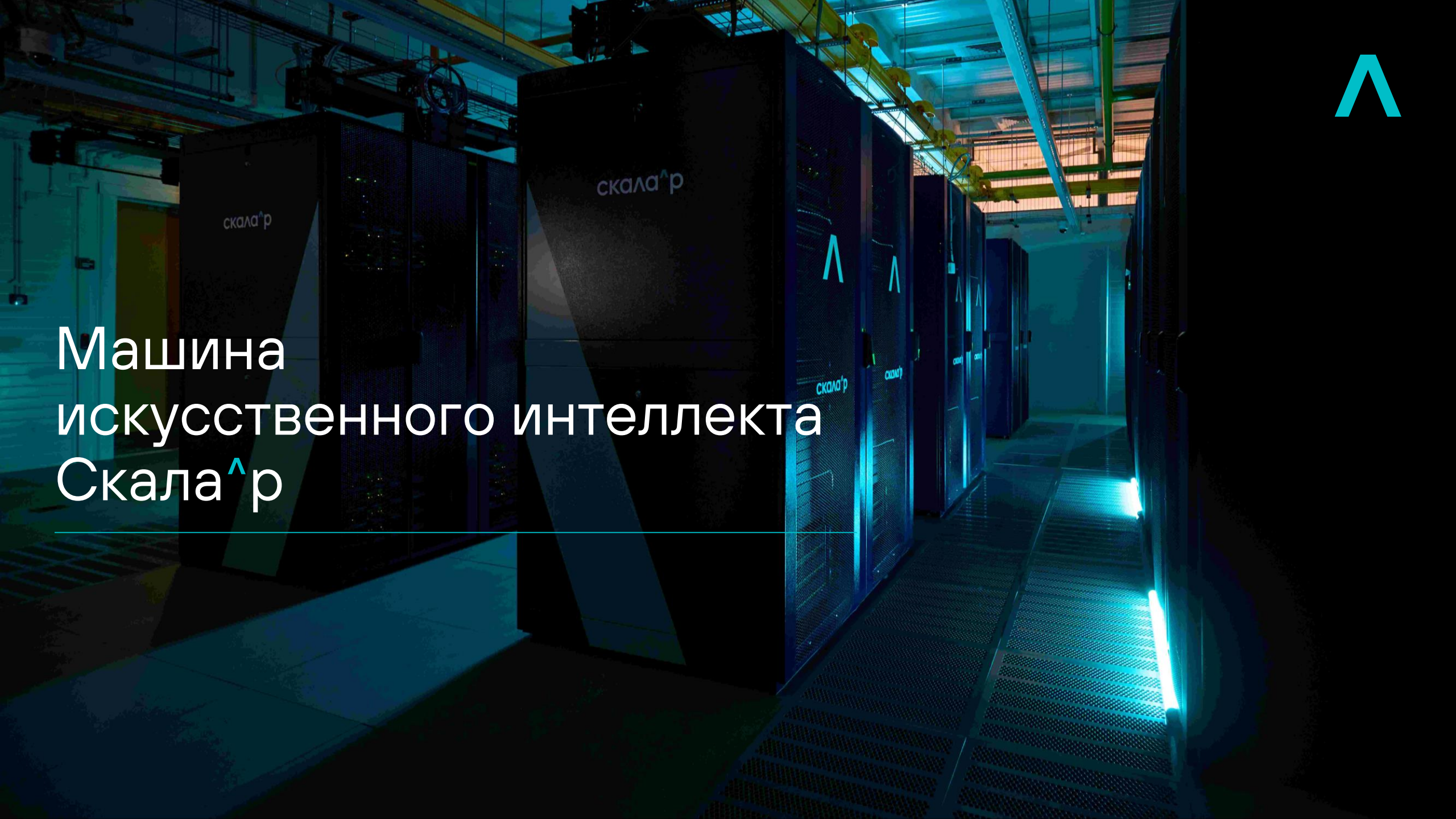
Инфраструктура безопасности Заказчика

- Системы обеспечения, мониторинга и поддержки ИБ
- SIEM

- Использование сертифицированных ОС и служебных СУБД специальные лицензии
- Сертифицированные версии ПО виртуализации
- Сертифицированные версии функционального ПО ПАК
- Управление доступом (встроенное ПО IAM опционально)
- Совместимость с наложенными средствами защиты (антивирус, АПМДЗ опционально)
- Сканирование на уязвимости на этапе разработки
- Повышение защищенности при производстве ПАК (hardening)

Результат

- Безопасность без ущерба для производительности и отказоустойчивости
- Соответствие требованиям регуляторов
- Снижение вероятности реализации киберриска и минимизация его воздействия
- **Предсказуемость и гарантированная совместимость**



Машина искусственного интеллекта Скала^р

Метрики Машины Скала^р МИИ

Типовой ПАК МИИ VRAM Nvidia

~10 Тбайт HBM3

~9 Тбайт HBM3e
VRAM Китай

~8 Тбайт HBM3

Сеть РФ

до 1,6 Тбайт/с Eth
между узлами с GPU
с поддержкой RDMA

RoCEv2 и GPUDirect

В каталоге более

80 LLM

разных версий
под разные задачи
и отсутствие ограничений
на запуск любых
инструментов AIOps

до 500 k8s pod'ов на узел

Размер LLM

235B

параметров в один рабочий узел без квантования

Равномерное деление GPU ресурсов MIG

или

неравномерное
деление GPU

Мировой опыт ИИ в архитектуре Машины

Nvidia создала NVIDIA DGX POD (DGX+Mellanox+NetAPP/DDN/DELL/HP+NGC+Inception), Huawei создала Atlas 900
и только в формате ПАК достигаются наилучшие показатели производительности

Полный контроль над стеком — ключ к устранению узких мест

Такой комплекс обеспечивает **полный контроль над всей технологической цепочкой**, что позволяет выявлять и устранять узкие места на стыках компонентов. Вертикальная интеграция аппаратуры, сети и софта дает возможность проводить сквозную оптимизацию, недоступную для разрозненных решений.

Целостность комплекса — основа надежности

Целостность комплекса обеспечивает беспрепятственный переход от разработки к промышленной эксплуатации. Единый каталог моделей и контейнеров устраняет «дрейф зависимостей» и **гарантирует воспроизводимость результатов**.

Единая система безопасности как результат полного контроля

Владение всем стеком ПАК позволяет выстроить единую, **сквозную систему информационной безопасности**, исключающую конфликты между разнородными компонентами. Полный контроль над кодом, конфигурациями и аппаратным обеспечением дает возможность проводить детальный аудит и проактивно выявлять уязвимости на всех уровнях.

Преимущества Машины Скала^рМИИ

Платформа данных как обязательная основа: качественные, доступные и валидируемые данные

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 МИИ

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

Внедрение практики непрерывного совершенствования

Концентрация на направлениях с измеримым экономическим эффектом

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

Реализация собственной ИИ-платформы, единая архитектура для разработки, внедрения и масштабирования ИИ-сервисов

Привязка всех ИИ-инициатив к бизнес-показателям и результатам KPI для каждого сервиса

* Показатели могут варьироваться в зависимости от задачи

Машина Скала^р МИИ

Выгоды бизнеса

Кратный рост прибыли

- Снижение FTE*, снижение стоимости операций, рост доходности и конверсии

Независимость от демографических провалов

- ИИ снимает зависимость от кадровых дефицитов и «демографических ям» — рост бизнеса больше не ограничен числом доступных специалистов

Экстерриториальность и гибкость

- Возможность для бизнеса работать в любом регионе, поддерживать переезд сотрудников и обеспечивать непрерывность операций без привязки к географии

Более сильные позиции в своей отрасли

- Отсутствие или недофинансирование ИИ-решений в составе бизнеса уже приводит к технологическому отставанию, росту себестоимости и потере доли рынка; в ближайшие годы это станет прямой причиной исчезновения слабых игроков

Более эффективное управление бизнес-процессами за счёт

- Тотальной цифровизации и интеллектуализации, E2E-описания** и пересборки ключевых процессов
- Оцифровки всех процессов и построения единой процессной архитектуры
- Обязательного мониторинга эффекта и контроля достижения результатов после внедрения

* **FTE (Full-Time Equivalent)** — эквивалента полной занятости в управлении персоналом и бухгалтерии
** Сквозное, от англ. E2E (End-to-End)

Разрешение рисков и сложностей с Машиной Скала^р МИИ

ПАК МИИ

Риск несовместимости ИИ-стека Новые подходы и продукты ИИ появляются ежедневно и в ходе реализации проекта будут регулярно меняться	Набор проверенных моделей и инструментов управления
Дефицит кадров Проекты ИИ зависят от ключевых сотрудников. Экспертов мало, они дороги и за ними идёт охота	Инструмент для развития собственных ИИ специалистов
Высокая стоимость инфраструктуры Оборудование дорогое, через год надо делать апгрейд, масштабирование зачастую неэффективно	Оптимизирован стек технологий для дальнейшего масштабирования
Риски безопасности Новые модели угроз и способы защиты	MLSecOps & TRiSM
Риски потери производительности «модель отлично работает на учебном стенде, но ее инференс в продакшене слишком медленный и дорогой»	Исследование всего аппаратного стека для ускорения вычислений и передачи данных

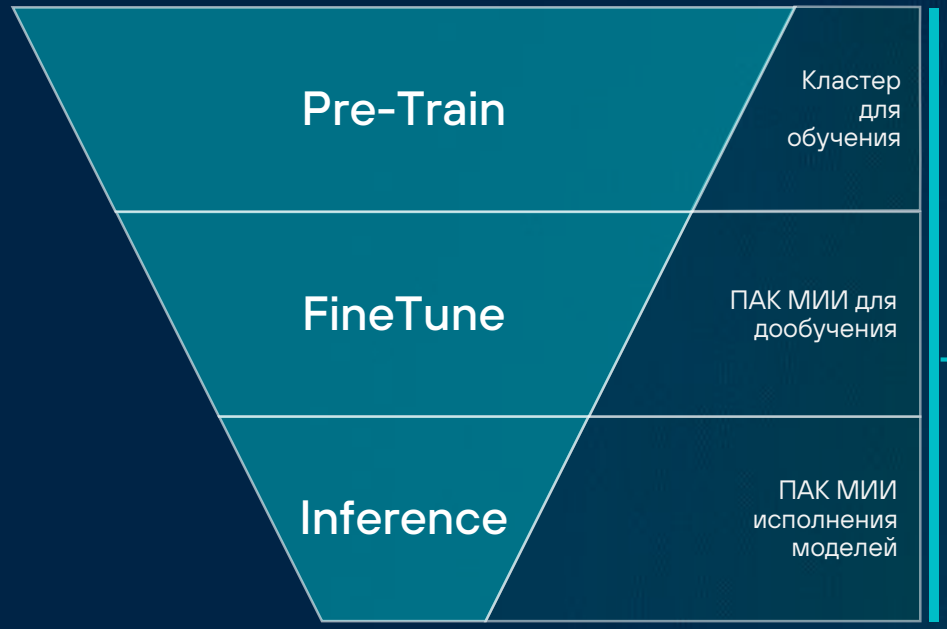
Улучшение показателей

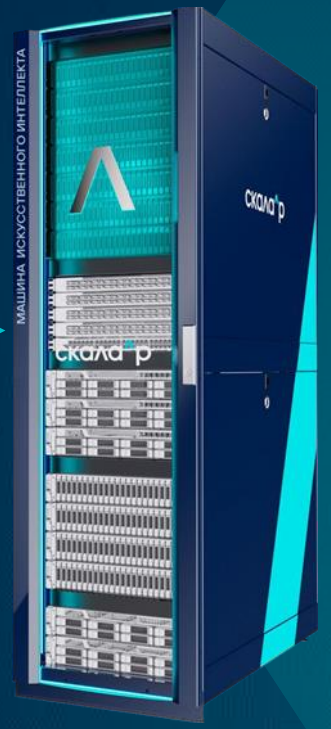
₽/Token ₽/Flops ₽/Tops

Машина Скала^р МИИ

Задачи

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





Аналог:
Huawei Atlas 900 Pod
NVIDIA DGX SuperPOD

- Cotype
- MLSecOps
- Yandex
- Just AI
- Смарт Платформа
- ValueAI
- GigaChat
- Llama
- DeepSeek
- другие ИИ решения

Машина искусственного интеллекта Скала^р МИИ

ПАК может исполнять модели LLM/SLM и ML, а также являться инструментом разработки и управления жизненным циклом ИИ решений

Доступные платформы MLOps и LLOps

n8n Dify скала^р в составе ПАК

Доступные Модели ИИ

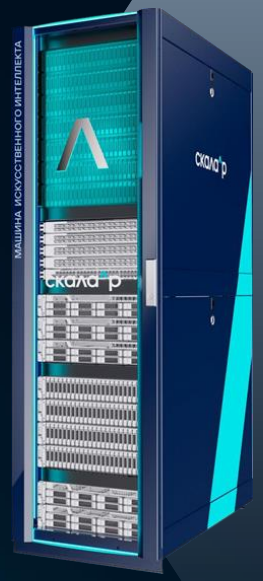
Языковые модели

- GigaChat
- YandexGPT
- DeepSeek
- Cotype

Модели машинного обучения

- Линейная регрессия
- Метод опорных векторов
- Дерево решений
- Модель случайного леса

DS платформа



ПАК состоит из трёх логических слоёв и стека технологий и решений

Спектр ИИ - ML оркестратор | Каталогизация ИИ контейнеров | Разработка ИИ-агентов

Постановка задач на запуск ИИ | Создание ИИ агентов | Интерфейс управления

Бенчмаркинг HW и LLM | OCR | Транскрибация

SW

Контейнерная инфраструктура | Наложенные средства ИБ

ОС

Драйверы | Прошивки | SDK

Скала^р Геном (мониторинг и управление ПАК)

MW

Аппаратное обеспечение

Вычисления: CPU | GPU | TPU

Сеть передачи данных: DPU | SNIC | RDMA; 10/25Гб/с | 100- 400Гб/с

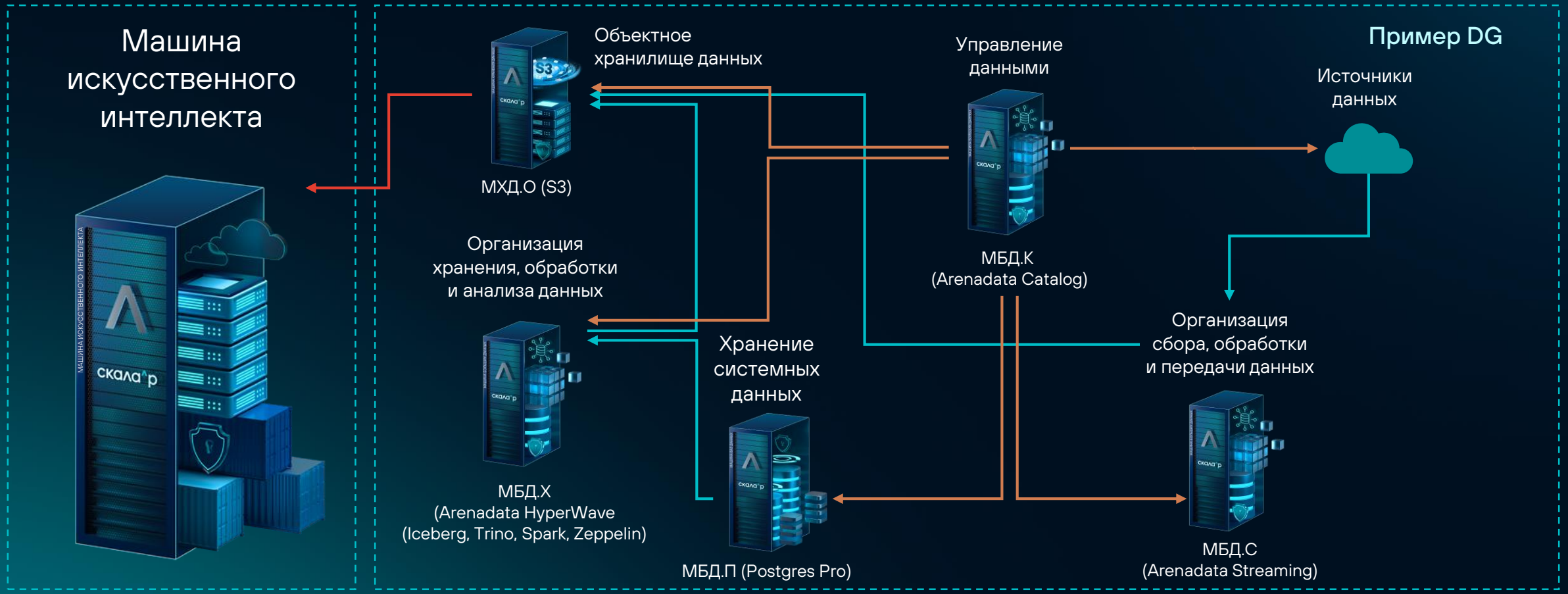
Хранение

Файловое | Блочное | S3

HW

Экосистема Скала^р для задач ИИ

продукты Скала^р максимально совместимы для единого решения комплексной задачи внедрения ИИ

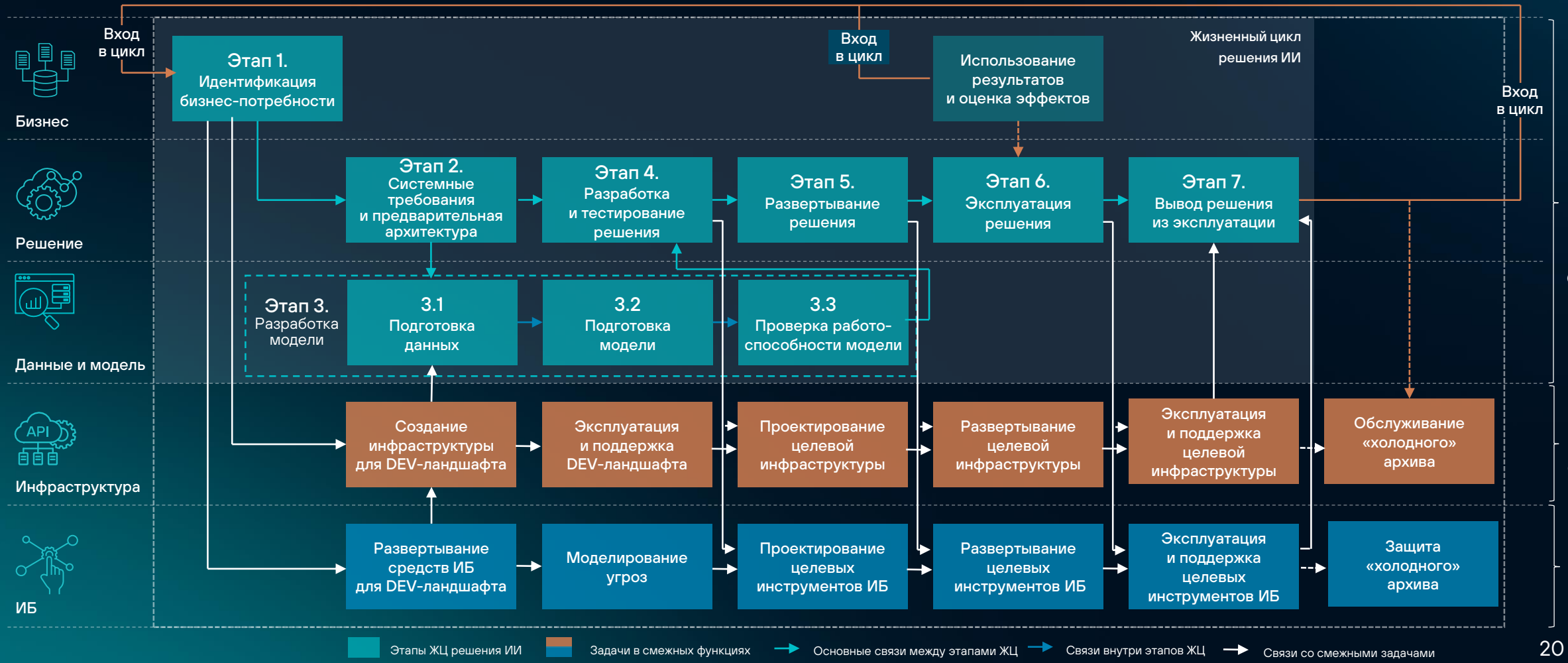


Экосистема Машин Скала^р позволяет минимизировать время и сложность внедрения ИИ в работу компаний и организаций

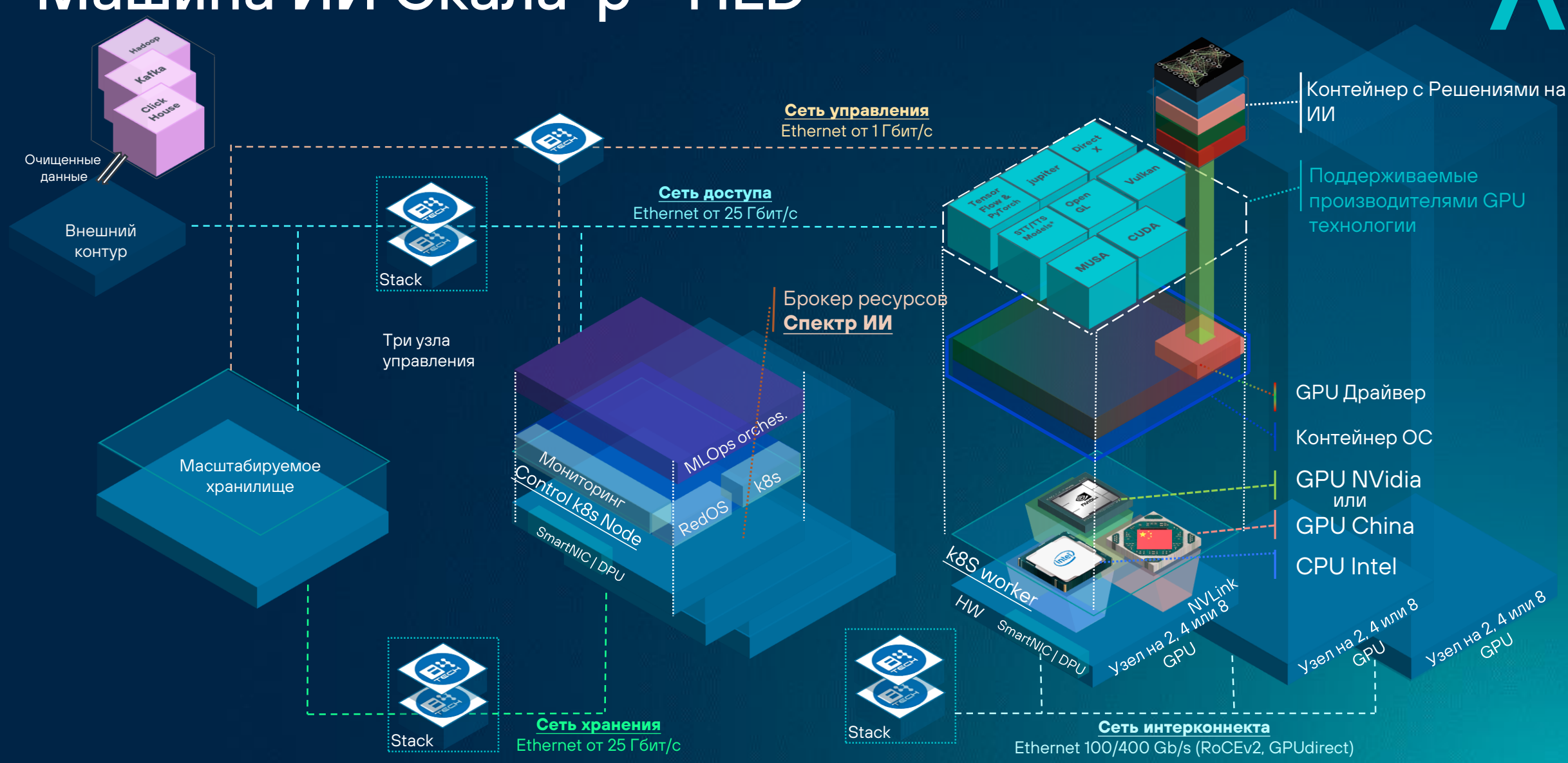
Жизненный цикл КИС с ИИ

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



Машина ИИ Скала^р - HLD



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-боты 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

Условные размеры Машины ИИ Скала^р

	Тестирование / Демо	Разработка / MVP	Продуктивный (корпоративный)	AI Factory (полный цикл)
Сеть	Сеть опционально	Сеть доступа Сеть управления	Сеть доступа/хранения Сеть управления Сеть интерконнекта	Сеть доступа Сеть интерконнекта Сеть хранения Сеть управления
Вычисления	GPU-узел на 4-8 GPU + NVLink или Сервер на 2 GPU	GPU-узлы на 4-8 GPU + NVLink	От 2-х GPU-узлов на 4-8 GPU + NVLink	До 16 GPU-узлов на 4-8 GPU + NVLink
Управление	Управление опционально	Сервер управления	2 сервера управления	3 сервера управления
Хранение	Хранение локально	Хранение 1) Локально 2) Внешний S3	Хранение 1) Локально 2) Внешний S3	Внешнее масштабируемое хранилище
	S* Инференс	M* Инференс	L* Инференс Дообучение Обучение	XL* Инференс Дообучение Обучение

* Все типоразмеры Машины ИИ, указанные на данном слайде, являются справочными (ознакомительными) и приведены для общего понимания концепции Изделия.
Данные типоразмеры Машины ИИ не являются окончательными коммерческими предложением (офертой) и не могут быть использованы
для составления договора поставки или иных юридически значимых документов без предварительного согласования и подтверждения в спецификации

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*

Машина Скала^р МИИ

Примеры исполняемых задач

- YandexGPT (LLM)
- RAG (Инструмент)
- Cotype (LLM)
- DeepSeek (LLM)
- ValueAI (Инструмент)
- Llama (LLM)
- GigaChat (LLM)
- другие ИИ решения

Вариант с применением GPU платформы с NVLink

(узел на 4-8 GPU + NVLink)

Вариант с применением типовых серверов 2RU

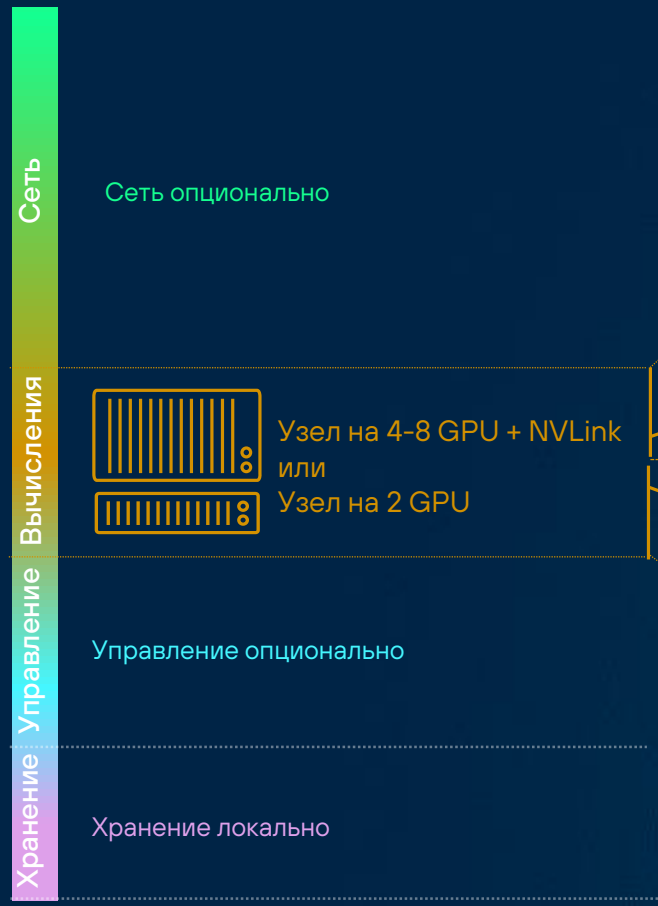
(узел на 2 GPU)

- 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

- 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 типоразмер S

Тестирование/Демо



Юниты	4 RU
Процессоры	2x Intel Xeon 5 Gen
RAM	до 4 Тбайт DDR5
PCIe слоты	Передняя панель: поддерживает максимум один слот PCIe 5.0 Задняя панель: поддерживает максимум 10 слотов PCIe 5.0 (до 8 видеокарт двойной ширины H100)
Энергопотребление	~3,6 кВт

Юниты	2 RU
Процессоры	2x Intel Xeon 5 Gen
RAM	до 8 Тбайт DDR5
PCIe слоты	до четырёх PCIe 5.0 x16 и до трёх PCIe 5.0 x8 (до 2 видеокарт двойной ширины H100)
Энергопотребление	~1,3 кВт

Используемые GPU NVIDIA	NVIDIA H100	NVIDIA H200	NVIDIA RTX 6000 Pro
Используемые GPU Азия	16GB GDDR	32GB GDDR	Аналог NVLink ограничено, PCIe 4.0 и PCIe 5.0

Инференс-узлы типоразмер M

Разработка / MVP

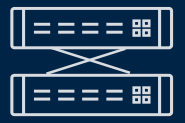
Сеть

Вычисления

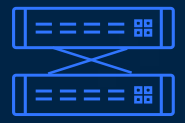
Управление

Хранение

Сеть доступа



Сеть управления



Узел на 4-8 GPU + NVLink

Узел управления

Хранение локально

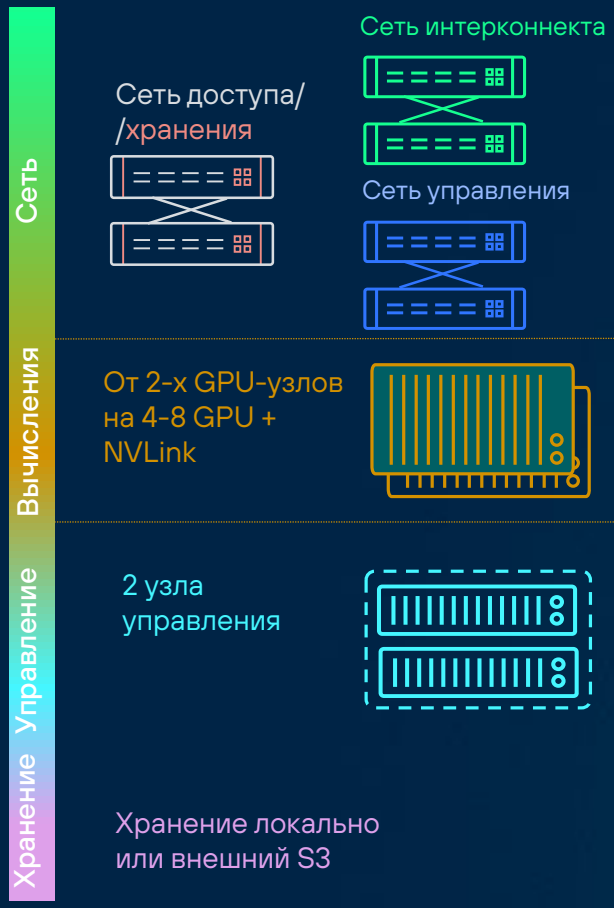
Юниты	4 RU
Процессоры	2x Intel Xeon 5 Gen
RAM	DDR5
PCIe слоты	до десяти PCIe 5.0
Энергопотребление	~4,8 кВт

Используемые GPU NVIDIA	NVIDIA H100	NVIDIA H200	NVIDIA RTX 6000 Pro

Используемые GPU Азия	16GB GDDR	32GB GDDR	Аналог NVLink ограничено, PCIe 4.0 и PCIe 5.0

Инференс-узлы типоразмер L

Продуктивный (корпоративный)



Юниты	4 RU
Процессоры	2x Intel Xeon 5 Gen
RAM	DDR5
PCIe слоты	до десяти PCIe 5.0
Энергопотребление	~4,8 кВт

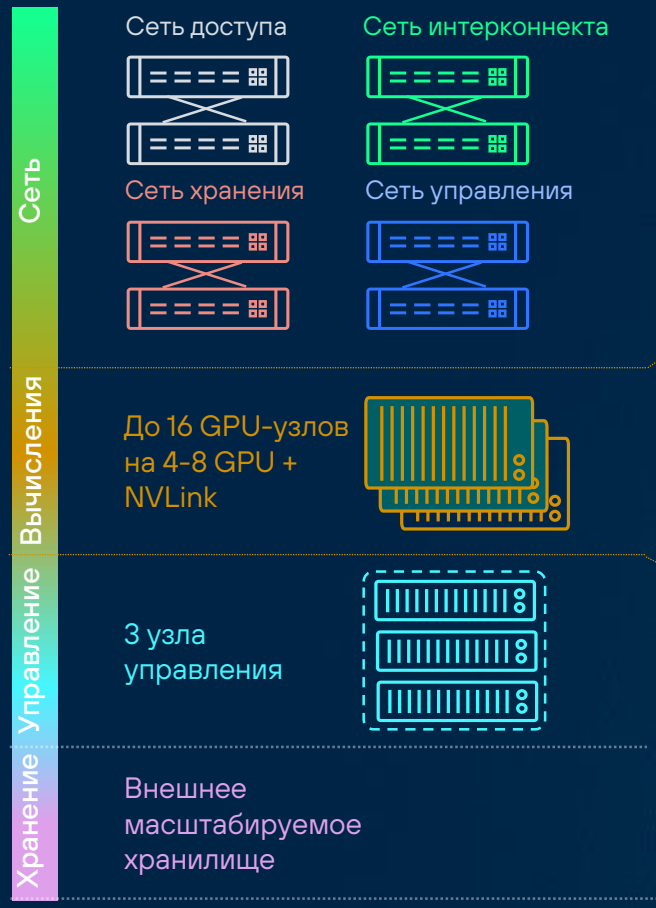
Используемые GPU NVIDIA	NVIDIA H100	NVIDIA H200	NVIDIA RTX 6000 Pro

Используемые GPU Азия	16GB GDDR	32GB GDDR	Аналог NVLink ограничено, PCIe 4.0 и PCIe 5.0

Инференс-узлы типоразмер XL

XL

AI Factory (полный цикл)



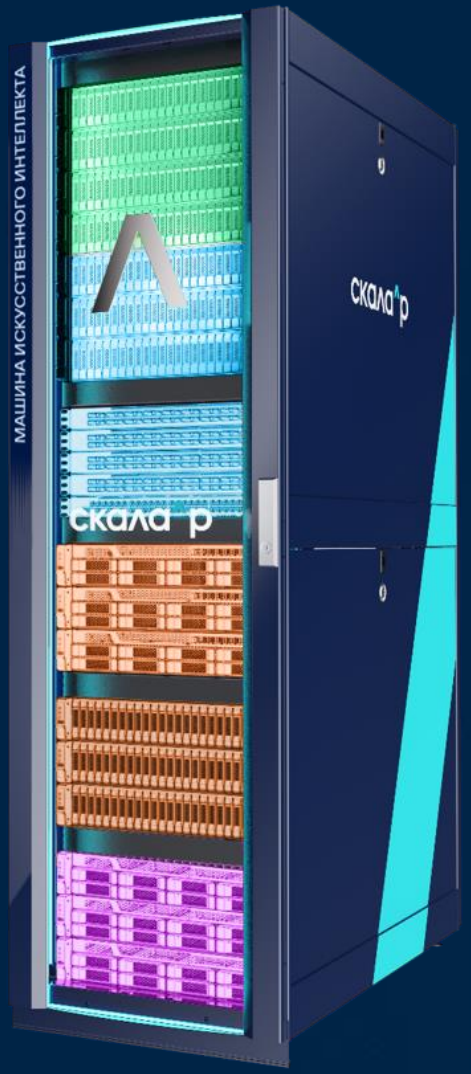
Юниты	4 RU
Процессоры	2x Intel Xeon 5 Gen
RAM	до 4 Тбайт DDR5
PCIe слоты	Передняя панель: поддерживает максимум один слот PCIe 5.0 Задняя панель: поддерживает максимум 10 слотов PCIe 5.0 (до 8 видеокарт двойной ширины H100)
Энергопотребление	~4,8 кВт

Используемые GPU NVIDIA	NVIDIA H100	NVIDIA H200	NVIDIA RTX 6000 Pro
Используемые GPU Азия	16GB GDDR	32GB GDDR	Аналог NVLink ограничено, PCIe 4.0 и PCIe 5.0

Машина Скала^р МИИ

XL

Модульная архитектура (на примере типоразмера XL)



Базовый модуль

Коммутация Машины

- 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)

Управление Машиной

- Три сервера для размещения управляющих компонентов 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

Функциональный модуль

- 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, до 4 Тбайт ОЗУ при оптимальной конфигурации памяти
- 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

Модуль хранения*

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

* В качестве системы хранения опционально могут использоваться модули хранения из состава Машин больших данных и/или Машин хранения данных производства ООО «СКАЛА-Р»

Машина Скала^р МИИ

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 ISW-7100)

Модуль управления Машиной МИИ (3 физических сервера)

ПО контейнеризации control plane и системные ВМ

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

ОС версия ФСТЭК — Сервер модуля управления 1

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

ОС версия ФСТЭК — Сервер модуля управления 2

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

ОС версия ФСТЭК — Сервер модуля управления 3

Модуль хранения (опция)

Отдельная система хранения данных

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

ПО контейнеризации версия ФСТЭК

ОС версия ФСТЭК

- Сервер модуля полезной нагрузки 1
- Сервер модуля полезной нагрузки 2
- Сервер модуля полезной нагрузки 3
- Сервер модуля полезной нагрузки 4

ОС версия ФСТЭК

- Сервер модуля полезной нагрузки 5
- Сервер модуля полезной нагрузки 6
- Сервер модуля полезной нагрузки 7
- Сервер модуля полезной нагрузки 8

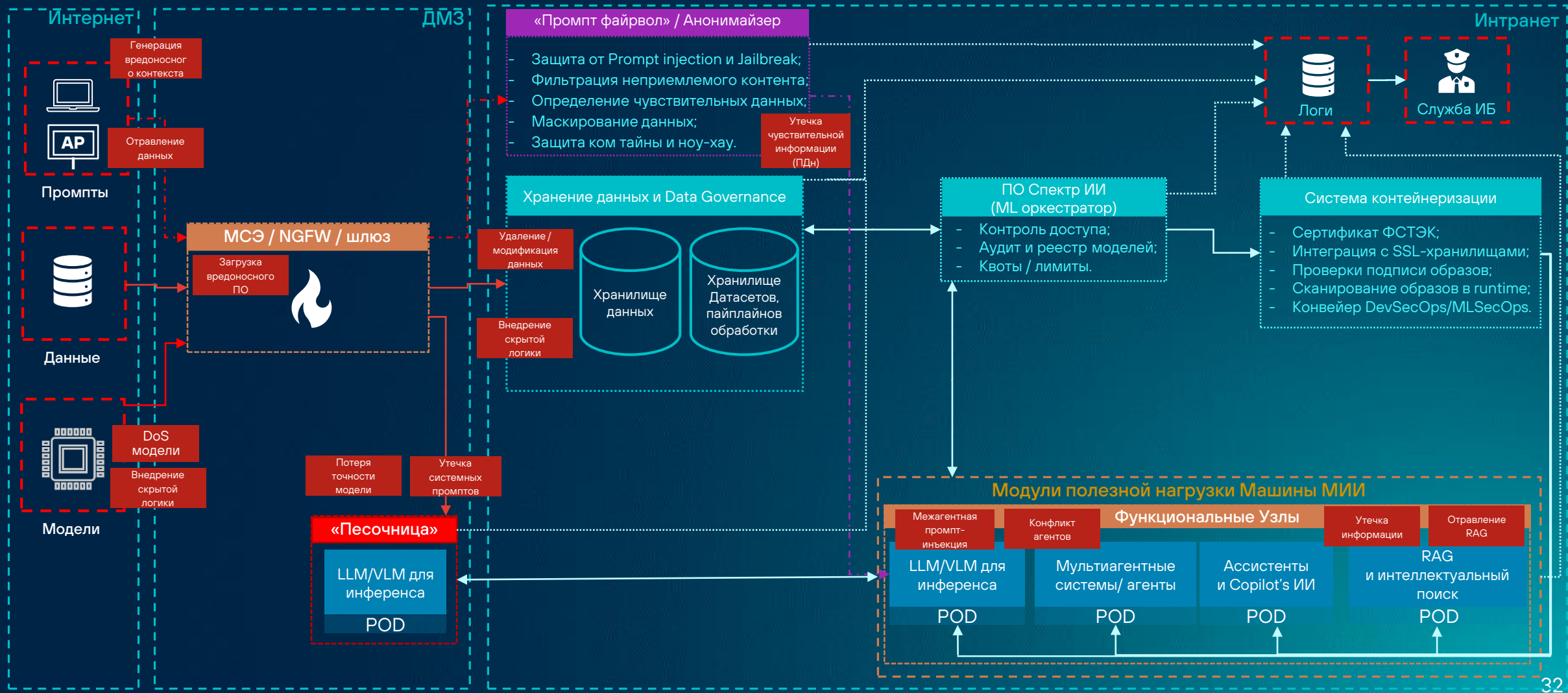
ОС версия ФСТЭК

- Сервер модуля полезной нагрузки 9
- Сервер модуля полезной нагрузки 10
- Сервер модуля полезной нагрузки 11
- Сервер модуля полезной нагрузки 12

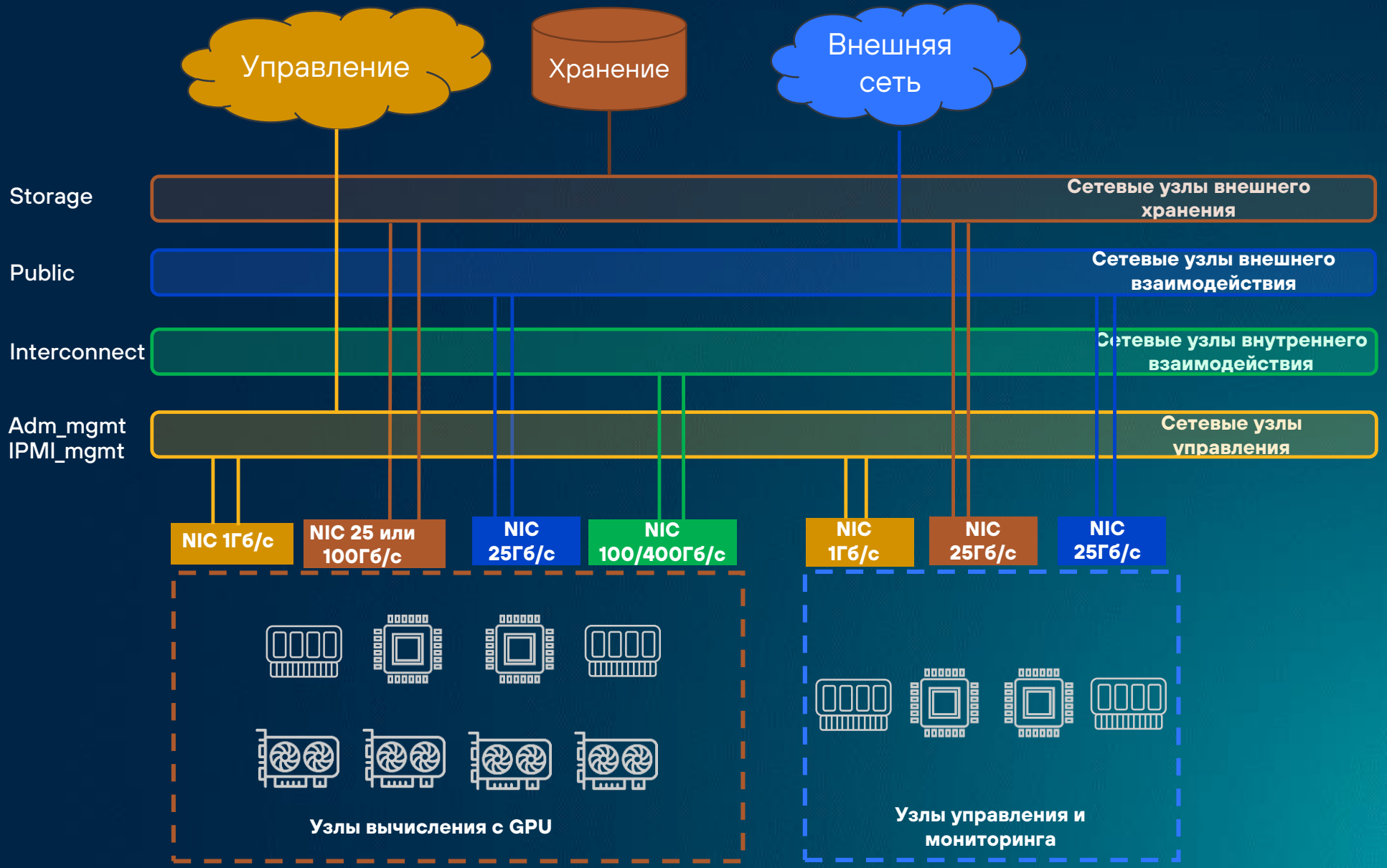
ОС версия ФСТЭК

- Сервер модуля полезной нагрузки 13
- Сервер модуля полезной нагрузки 14
- Сервер модуля полезной нагрузки 15
- Сервер модуля полезной нагрузки 16

Таксономия угроз ИБ для ИИ по международным фреймворкам



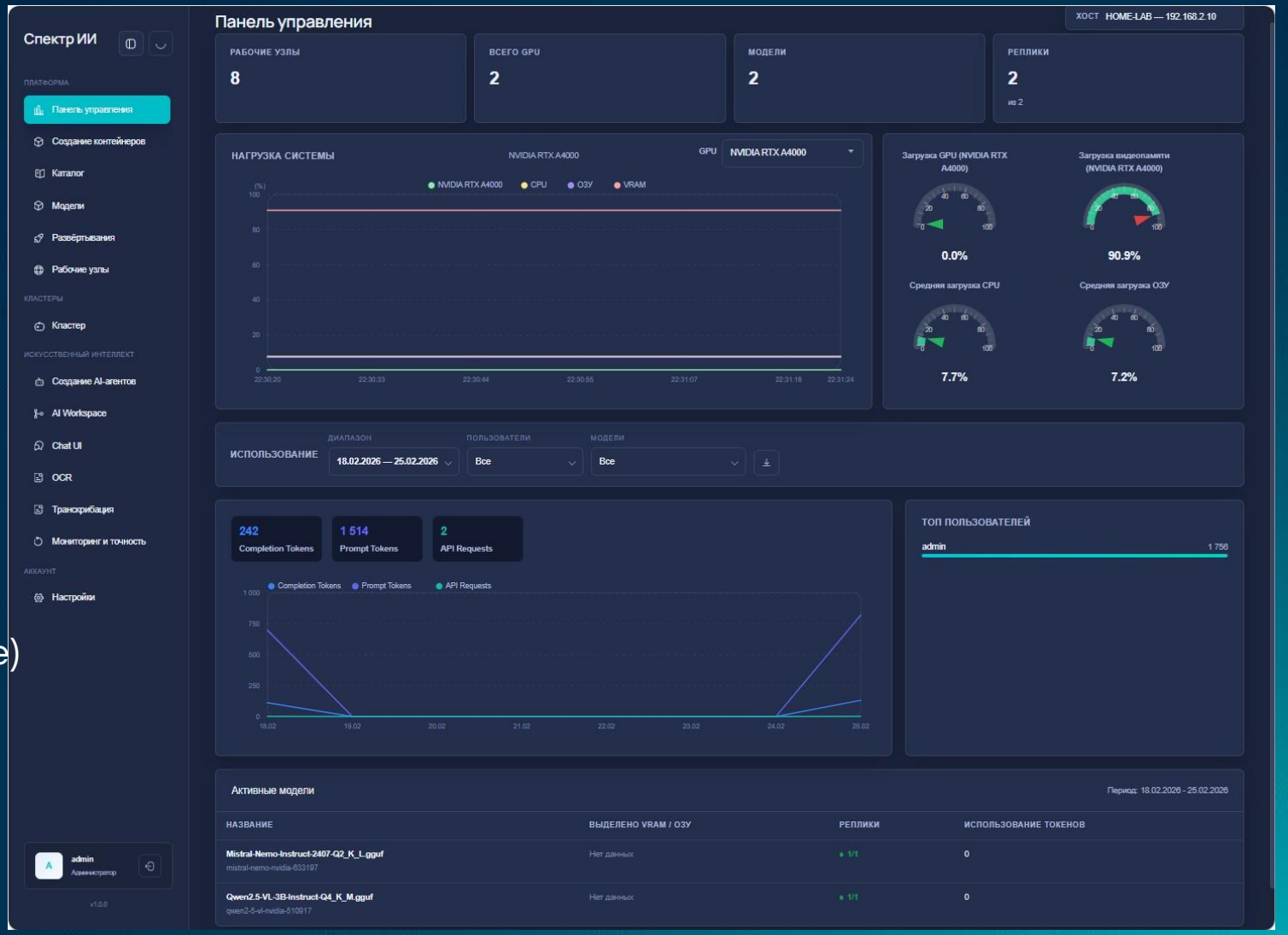
Типовая схема соединений и Vlan Машины МИИ



Спектр ИИ Скала^р

Функционал

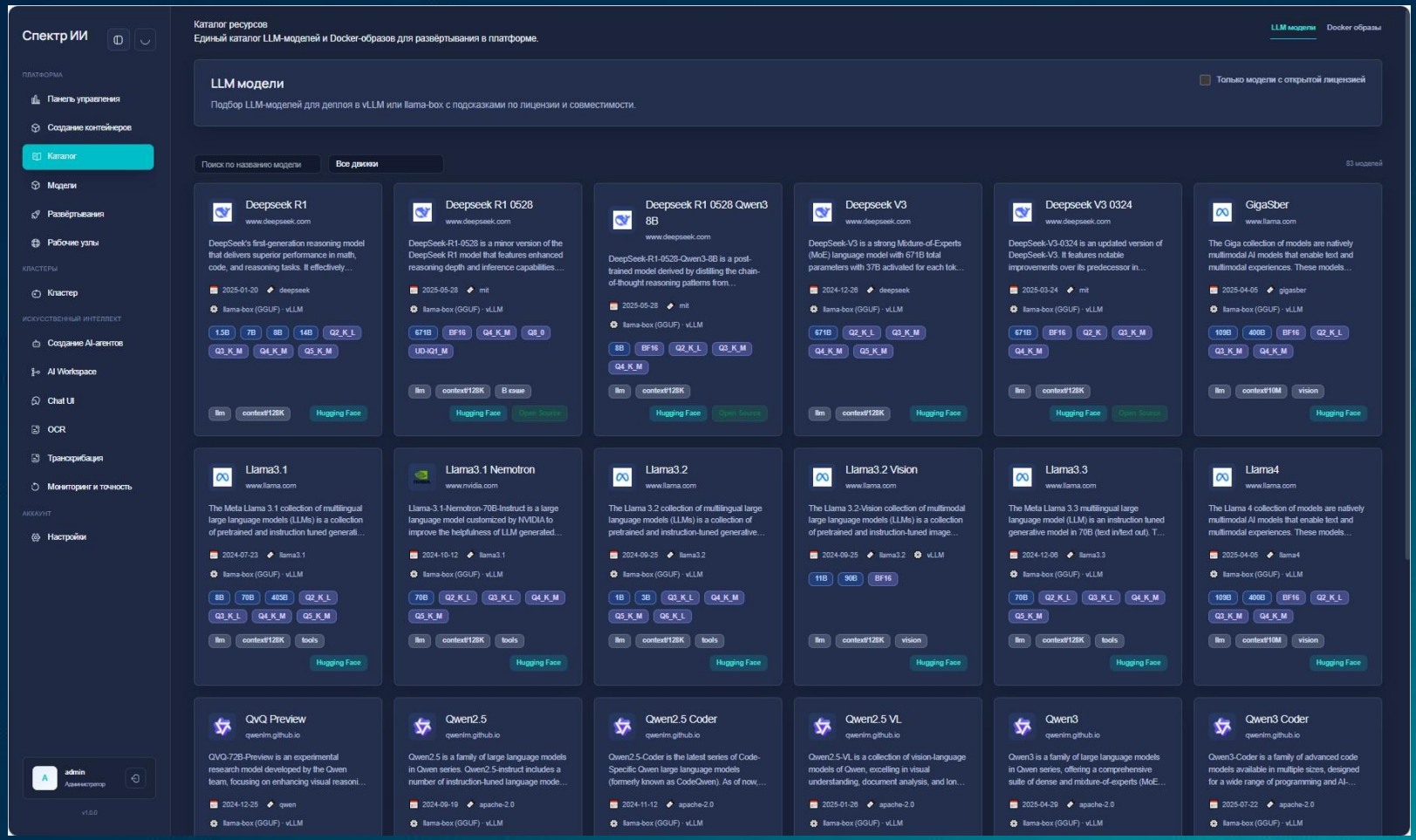
- управление ресурсами GPU платформы
- мониторинг нагрузки
- метрики производительности
- быстрое развертывание ИИ-инструментов, моделей (ML/LLM), сред разработки с применением технологий контейнеризации
- настройка применения моделей (ML/LLM)
- мониторинг и отчетность использования моделей (журналирование)
- набор инструментов для задач ИИ: распознавание текста, транскрибация, работа с AI-агентами



Спектр ИИ Скала^р

Каталог ресурсов

- **Каталог ресурсов Спектр ИИ** может содержать модели, загружаемые как с внешних ресурсов (Hugging Face и др.), так и внутреннего репозитория для дальнейшего быстрого их запуска, управления их жизненных циклом и мониторинга их использования (статус, путь хранения, размер, время создания, источник загрузки, используемый рабочий узел для их развертывания)



Спектр ИИ Скала^р

Основные разделы интерфейса

Интерфейс Спектр ИИ включает основные разделы по набору функциональных возможностей

- Раздел **Платформа** объединяет функции для мониторинга и управления инфраструктурой и инструментами для задач ИИ, использование технологий контейнеризации

- Раздел **Искусственный интеллект** содержит наиболее востребованный среди пользователей набор инструментов для задач ИИ: создание AI-агентов, транскрибация видео/аудио материалов, распознавание текста сканированных копий документов

- Раздел **Аккаунт** предоставляет возможность создания и настройки учетных записей с распределением доступов в зависимости от ролевой модели, прав пользователей/групп, а также настройки управления списком хостов и их параметрами

- Раздел **Кластеры** предназначен для пользователей Kubernetes, позволяющий с помощью Мастера первичной настройки добавить уже существующий кластер или создать кластер с помощью Мастера развертывания, выбрав один из сценариев

Спасибо за внимание!

www.skala-r.ru

Документ актуализирован 06.03.2026