



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 виртуализации Скала^р МВ

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

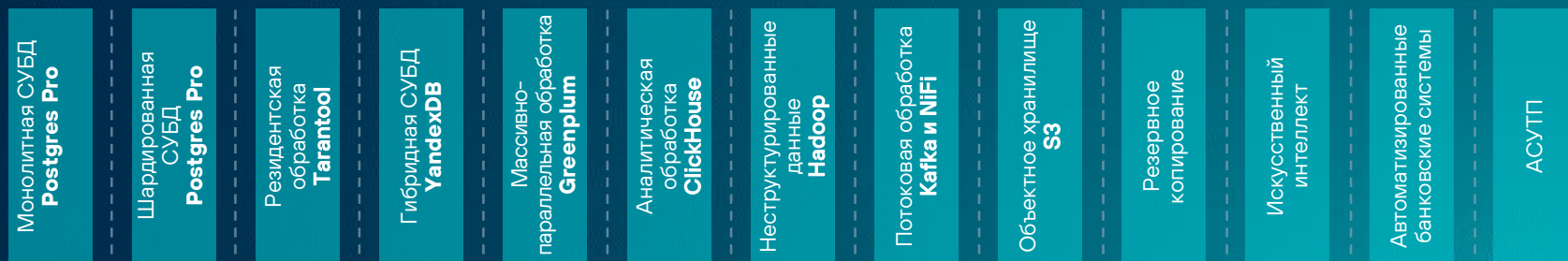
Модульная платформа

Динамическая инфраструктура



Динамическая инфраструктура

Инфраструктура управления данными



Транзакционная обработка

Большие данные

Специализир. решения

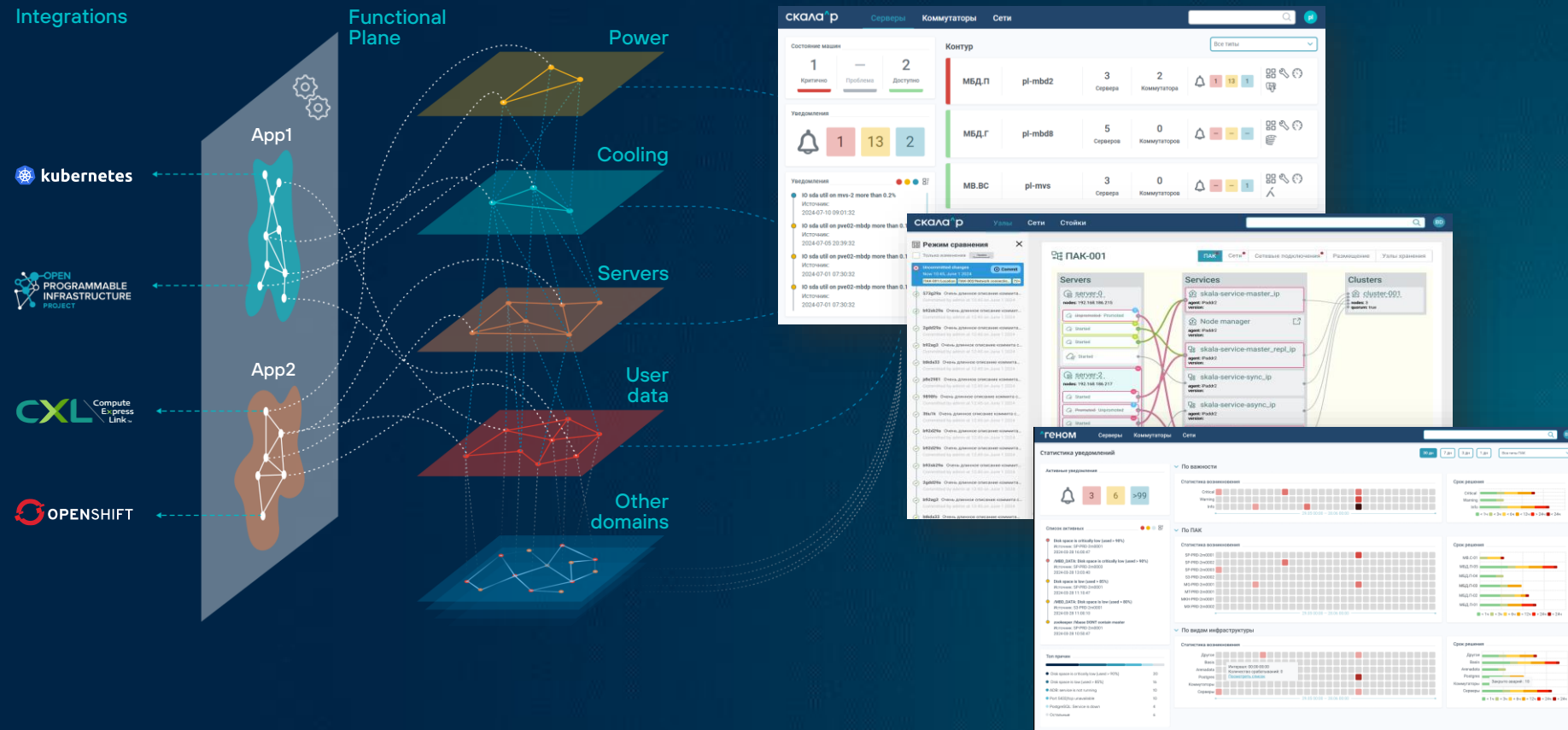
Отраслевые решения

Перспективная 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[^]p – Secure by Design



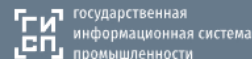
ПАК Скала[^]р в Реестрах РФ



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



Производительность

x2[↑]

чем решения, использующие сопоставимые аппаратные средства за счет оптимизации ввода-вывода 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

Единая техническая поддержка

Сопровождение оборудования и программного обеспечения всех компонентов Машин

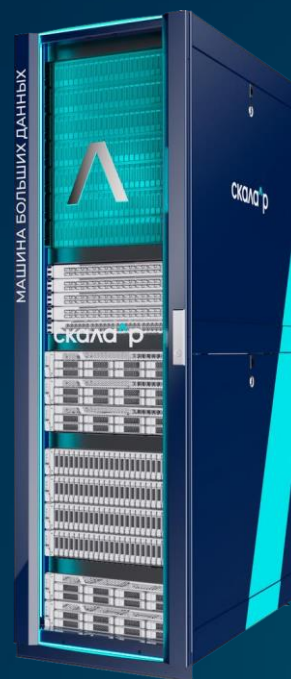
Экономия до 90%
на проектировании и внедрении

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

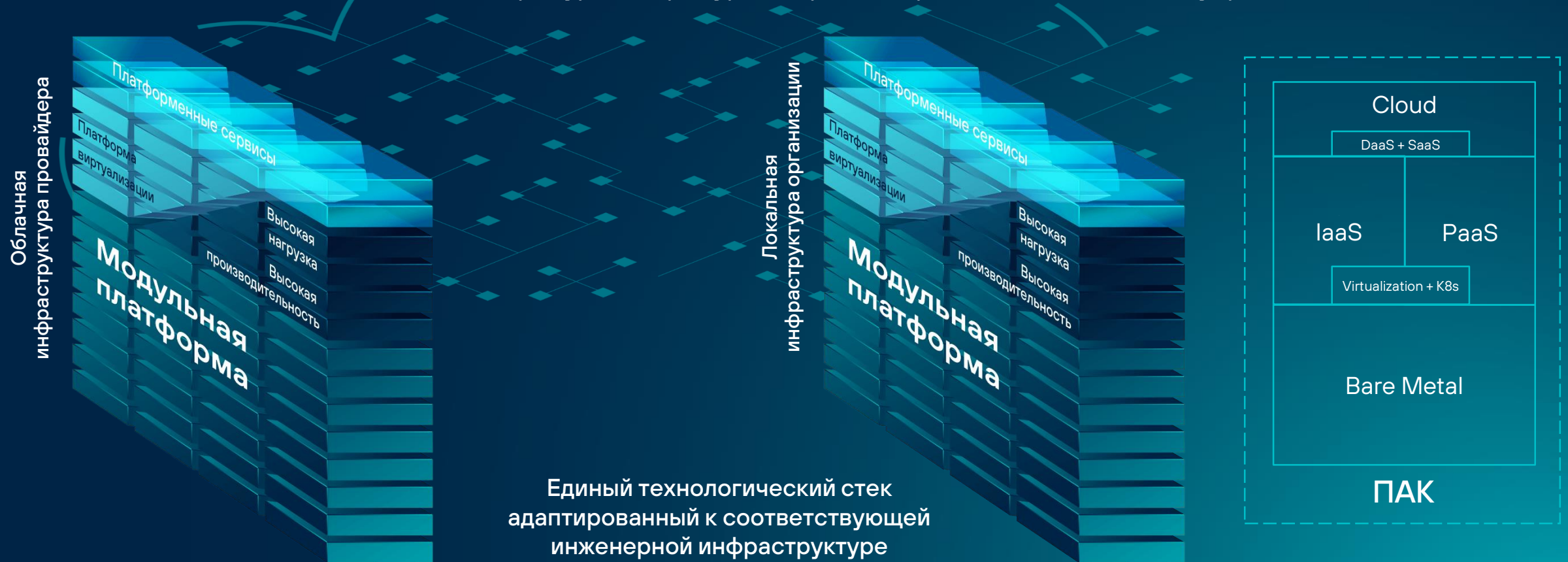
ПАК — Программно-аппаратный комплекс и модули платформы — включены в Единый реестр российской радиоэлектронной продукции и реестр Минцифры

Продуктово-технологическая концепция Скала^р

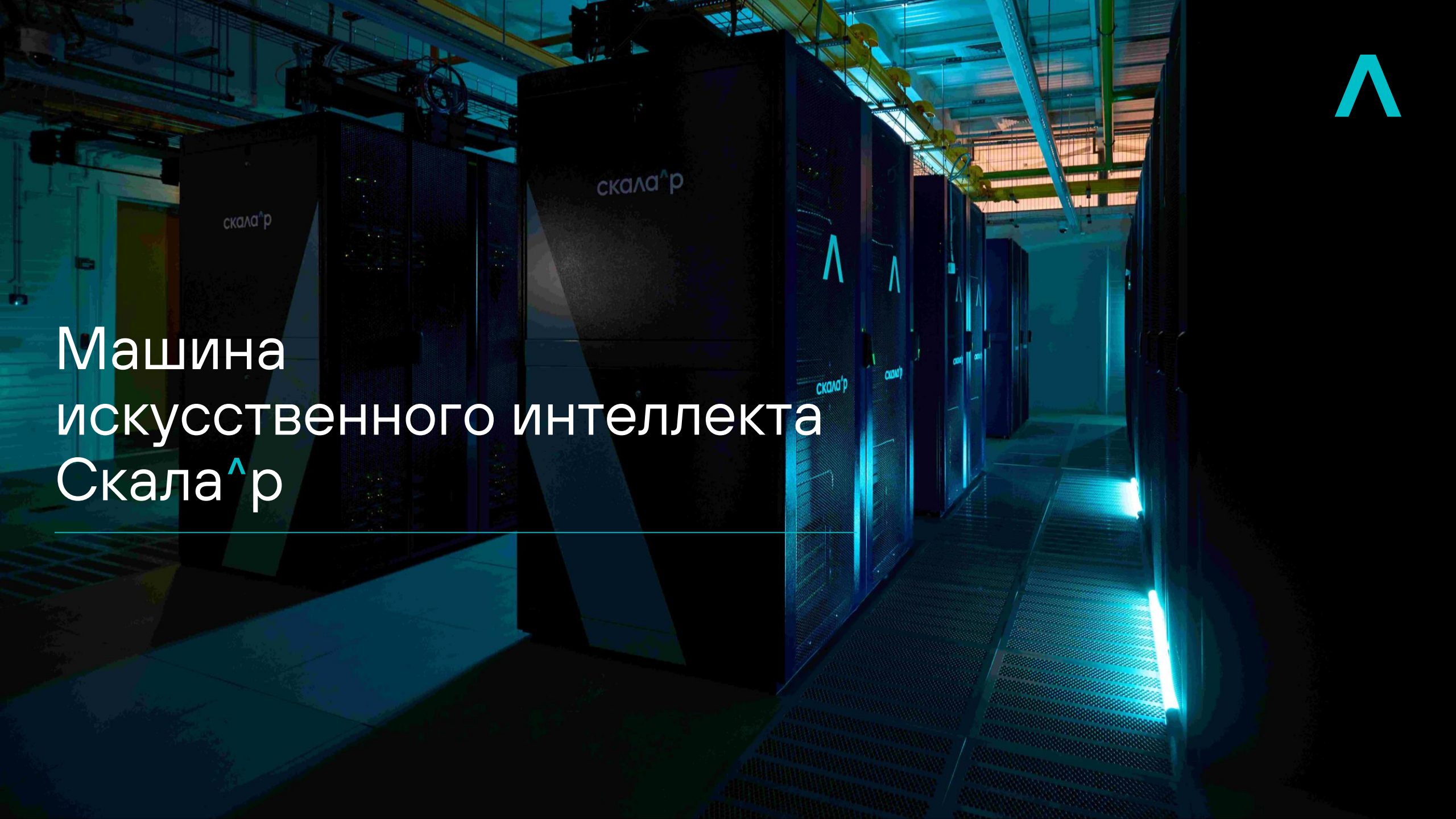
Миграция крупнейших организаций в распределенное облако

Построение локальных модульных инфраструктур с облачной системой управления от провайдера

Совместное использование локальных ресурсов и ресурсов провайдера из единой консоли управления



Если крупные корпоративные Заказчики не идут в облако провайдера, то облако провайдера должно прийти к ним



Машина искусственного интеллекта Скала^р

Машина искусственного интеллекта Скала^р



ИИ решения



Cotype



GigaChat



ChatGPT



Llama



Сайбокс



DeepSeek



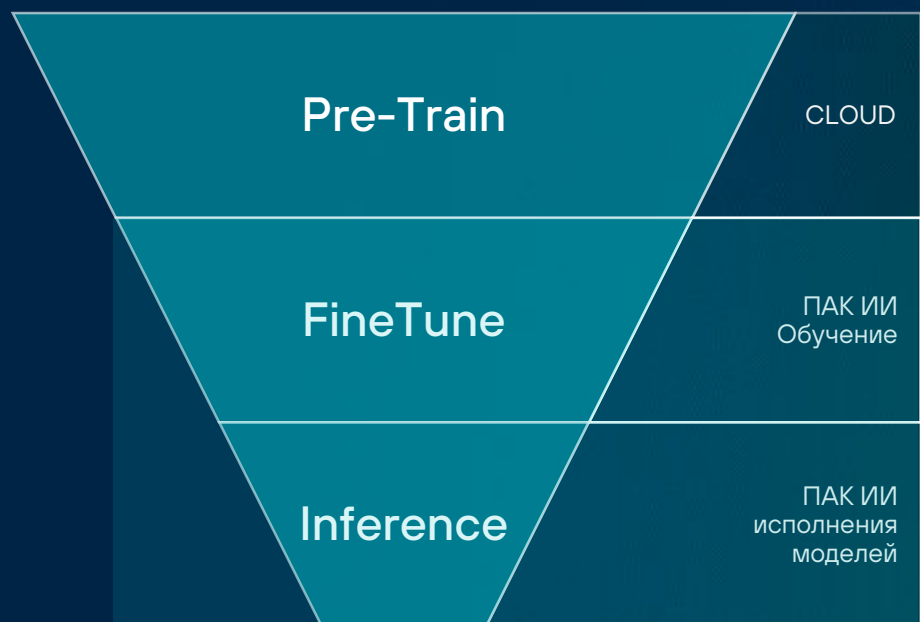
YandexGPT 5



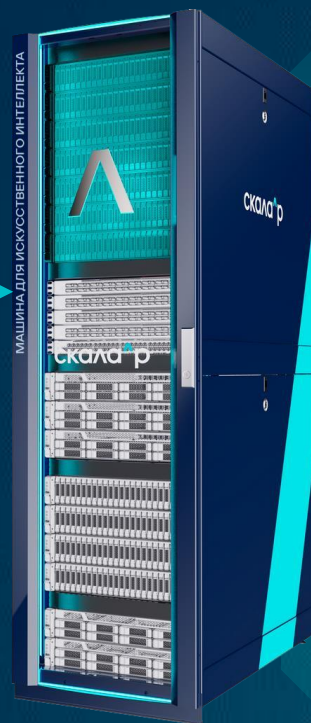
Смарт
Платформа



Другие
ИИ решения



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



Аналог

Huawei Atlas 900 Pod / NVIDIA DGX SuperPOD



Состав Машины ИИ Скала^Λр



Базовый модуль
и узлы управления
+ узлы коммутации



Модуль
вычисления
для задач:

- Обучения моделей
- Разработки ИИ решений

и/или



Модуль
вычисления
для задач
исполнения
(inference)



ПАК ИИ
(Fine tune)



ПАК ИИ
(Inference)

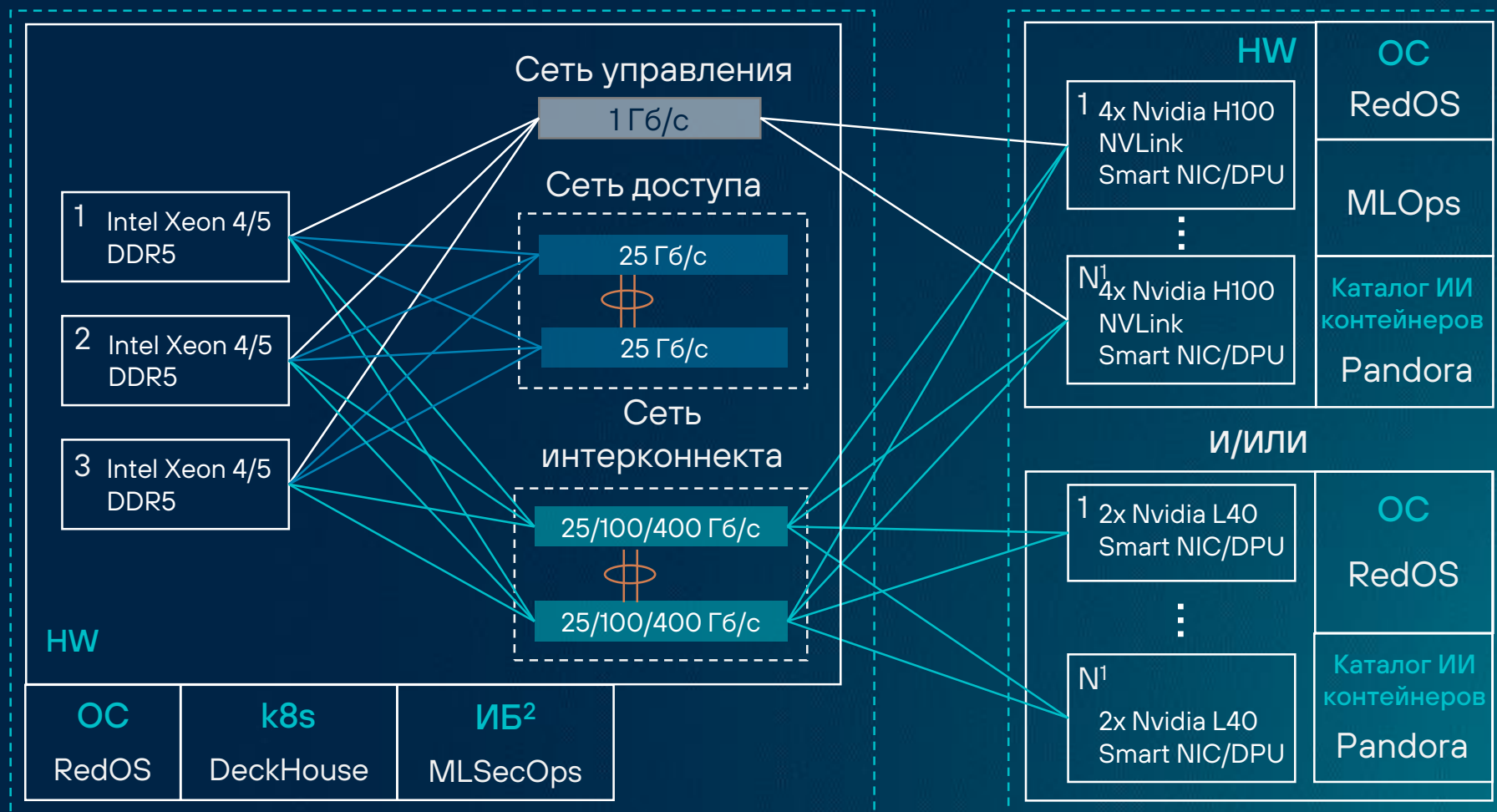


Состав Модулей Машины ИИ Скала[^]р



Базовый модуль и узлы управления
+ узлы коммутации

Модуль вычисления
и рабочие узлы



¹ Зависит от нагрузки

² Опционально

Сценарии применения Машины ИИ Скала^р



подтверждена работа и подтверждены характеристики, сопоставимые с облачными показателями, для различных ML платформ и моделей

Платформы MLOps и LLOps



Автоматизация



Контроль качества



Конвейер разработки



Мониторинг

Модели ИИ

Языковые модели



Llama



ValueAI

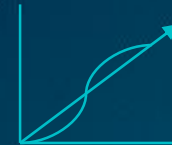


DeepSeek



Cotype

Модели машинного обучения



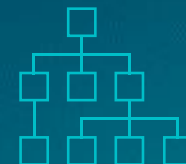
Линейная регрессия



Метод опорных векторов



Дерево решений



Модель случайного леса

Гибкая унифицированная архитектура в соответствии с современными отраслевыми стандартами, позволяет использовать её с доступными на российском рынке приложениями такими как:



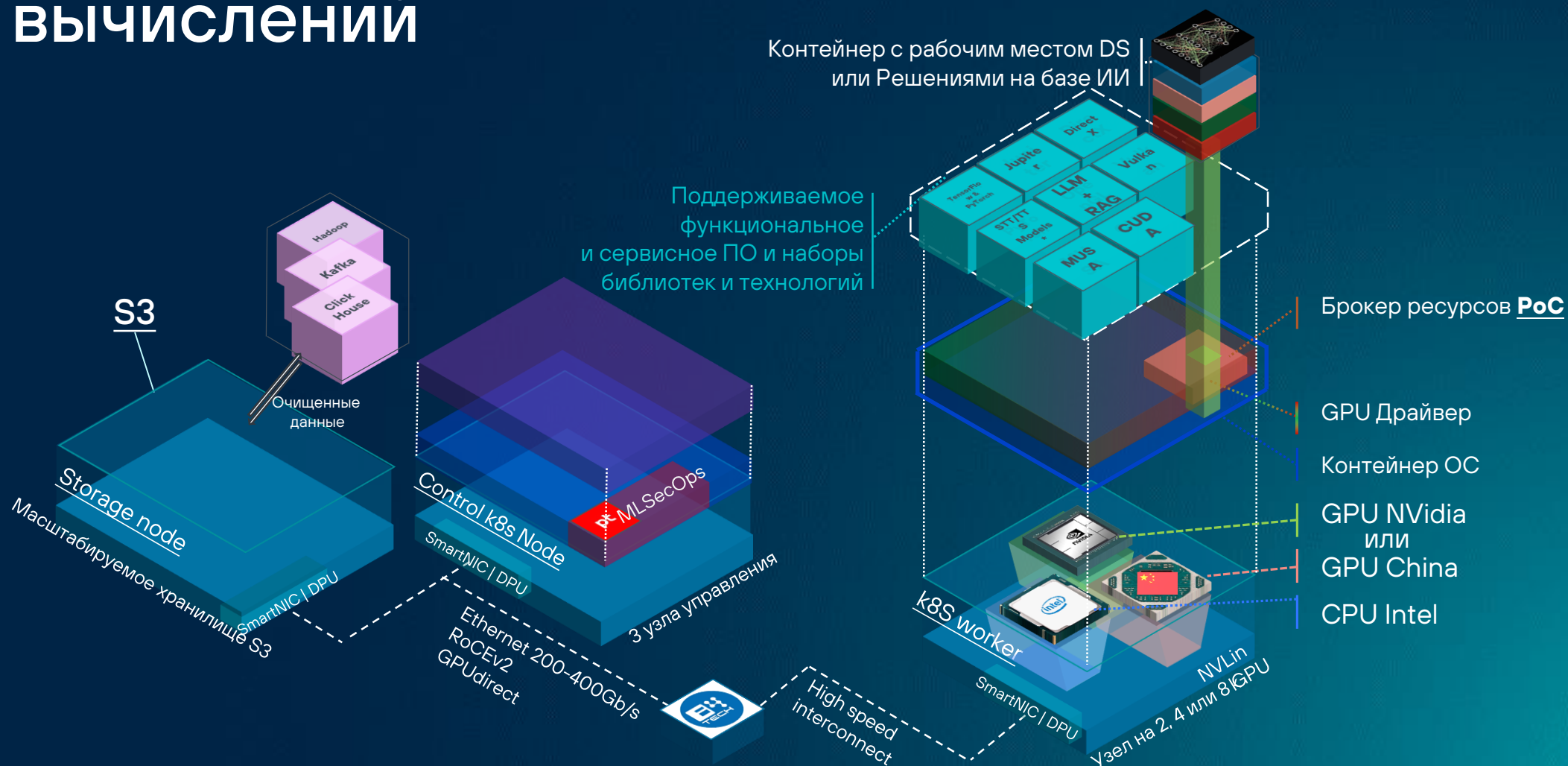
GigaChat



YandexGPT

и другими

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



Метрики по ПАКу ИИ



Производительность

≥ 6 Pflops

на один рабочий узел (TF32)

≥ 400 Tflops

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

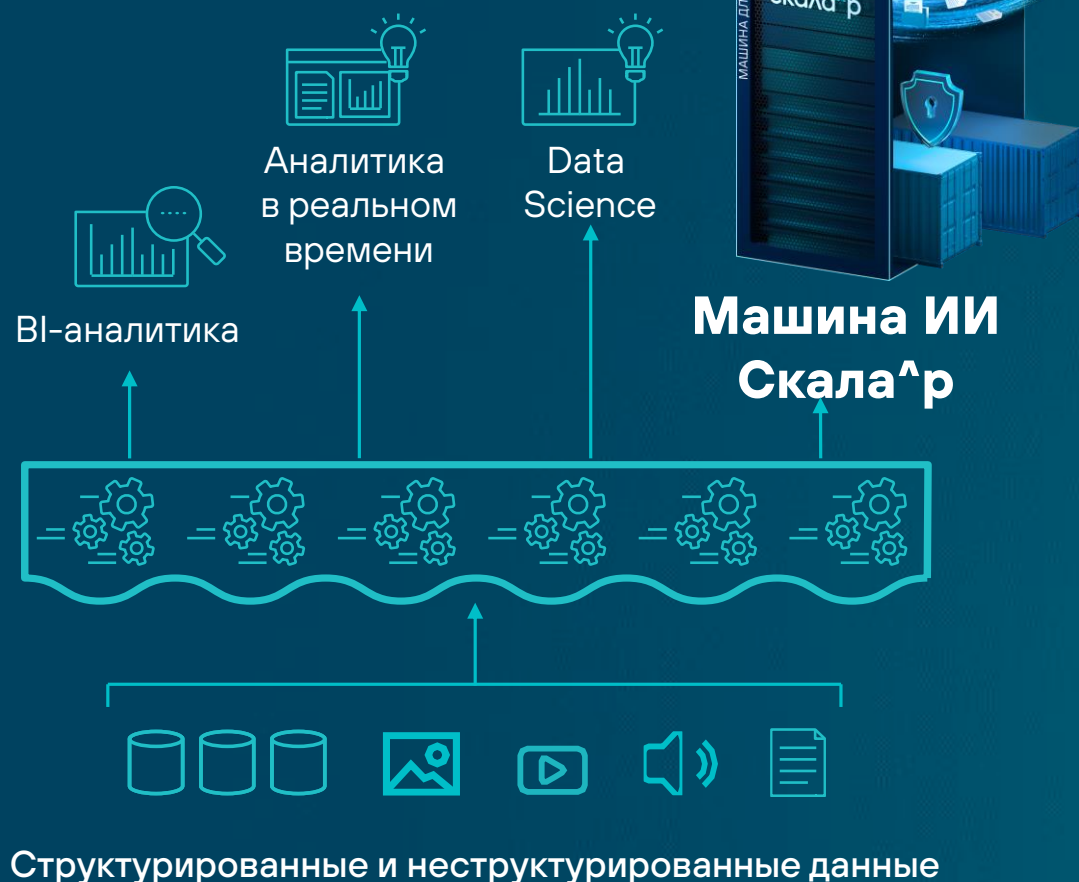
в 2–5 раз

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

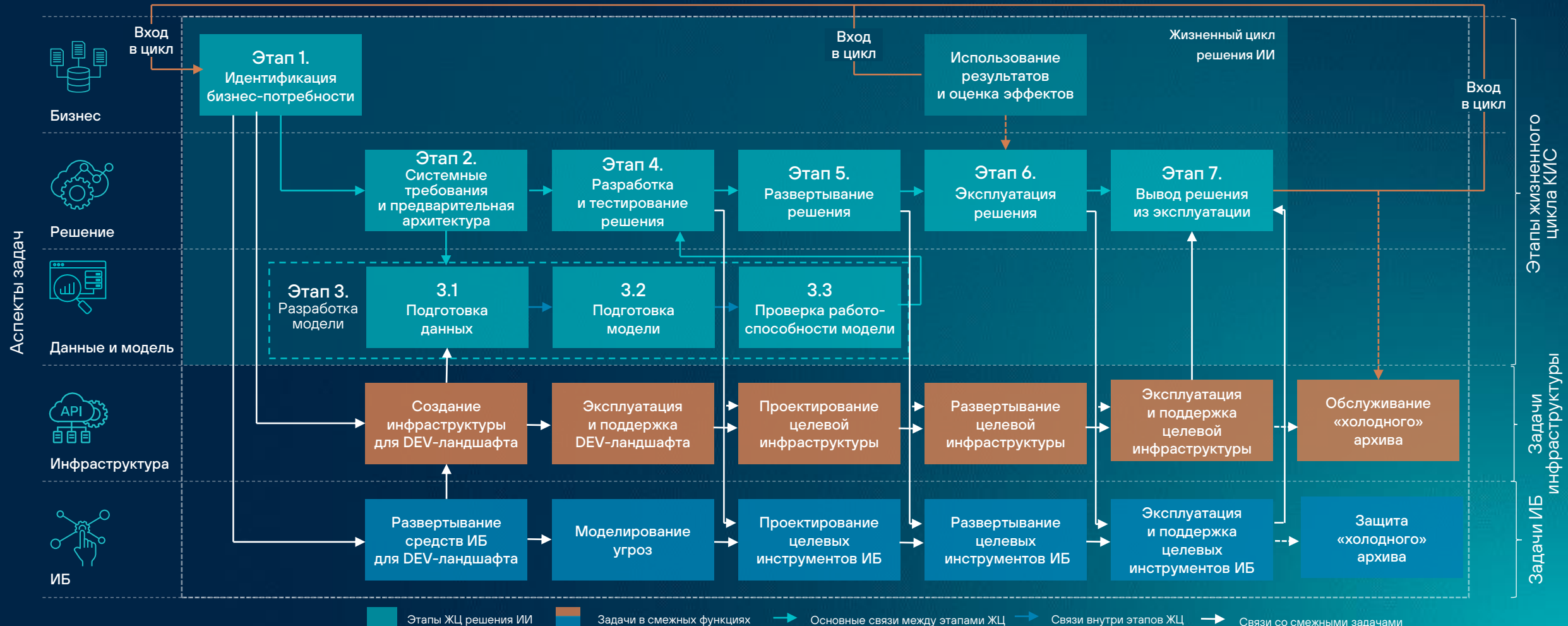
Система скрининга соискателей на соответствие требованиям позиции (вакансии). Построена на основе технологий NLP и применения адаптированных языковых моделей LLM

* Типовые задачи для инфраструктуры Машины ИИ Скала^p

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





www.skala-r.ru