deckhouse kept

От пилота до прода

Что может остановить
ИИ-агента?

Кто ведёт вебинар

Александр Подмосковный

Расскажет об инфраструктуре, GPU и сервисах инференса

Роман Малюга

Расскажет об агентах, инструментах и прикладном сценарии

deckhouse kept



Александр Подмосковный

Флант / Deckhouse Platform

Директор продуктового направления ML/AI

aleksandr.podmoskovniy@flant.com



Роман Малюга

Kept / KeptStore

Партнёр, финансовая трансформация и цифровые решения

RMalyuga@kept.ru

О вендоре Deckhouse

deckhouse

18+

лет работы
с Open Source

С 2017

года используем
Kubernetes в production

№ 1

контрибьютор в проекты
CNCF из России

500+

сотрудников

>260

компаний-пользователей

№ 1

в рейтинге вендоров систем мониторинга
и управления ИТ-инфраструктурой*

Реестр
российского ПО

Лицензии и сертификат
ФСТЭК России

АРПП
«Отечественный софт»

* Рейтинг «Российский рынок систем мониторинга и управления ИТ-инфраструктурой», TAdviser, 2026

Kept — одна из крупнейших аудиторско-консалтинговых фирм в России и Беларуси

79 из 100

крупнейших российских компаний являются нашими клиентами*

11

офисов работают в крупнейших городах России и Беларуси

3400

специалистов

35

лет работаем в России

* Рейтинг крупнейших по выручке компаний России — РБК 500, 2025 г. Анализ клиентской базы за 2025 г.

План вебинара

deckhouse kept

01 Зачем бизнесу агенты: пример с документами

02 Почему Kept и Deckhouse работают вместе

03 Инференс и управление GPU

04 Выполнение заданий и сопровождение

05 Экономика и выводы

На что уходит время при работе с документами

deckhouse kept

Сравниваем время, качество и стоимость обработки

Ручная работа

- Чтение документов
- Ручной перенос в 1С
- Сверка и исправления

Цель автоматизации

- **Меньше ручной обработки**
- Повторяющиеся операции с проверяемым итогом
- В действующей 1С

Как агент обрабатывает документ

Пример сценария Kept:
приёмка ТМЦ в 1С

deckhouse kept

01 **Распознавание**

Текст документа
OCR

02 **Извлечение**

Поля документа
Языковая модель

03 **Действие**

Операция в системе
API или GUI

04 **Сверка**

Документ и данные 1С

Демонстрация Kept

deckhouse kept

Приёмка ТМЦ в 1С

Загрузка → обработка → результат в 1С
Модели, GPU и виртуальное рабочее место

Почему Kept и Deckhouse работают вместе

deckhouse kept

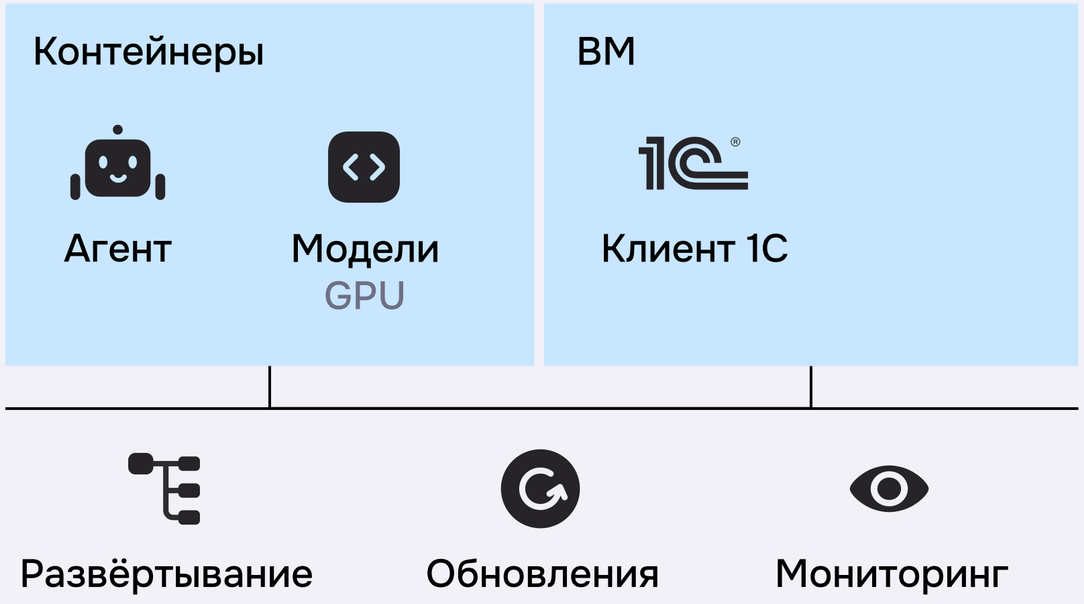
KeptStore

Прикладная логика

- Обработка документов
- Управление заданиями
- Инструменты 1С

Deckhouse Platform

Единая среда



Как агент работает с моделями и 1С

deckhouse kept



Как будем запускать модели из модулей Deckhouse

deckhouse kept

Модуль модели

Версия модуля

- Веса модели
- Движок инференса

Проверенные конфигурации

Запуск

Сервис инференса

- Движок инференса
- Загруженная модель

GPU компании

Запрос по API

Ответ

Приложения

- ИИ-агент
- Поиск по документам

Управление GPU

deckhouse kept

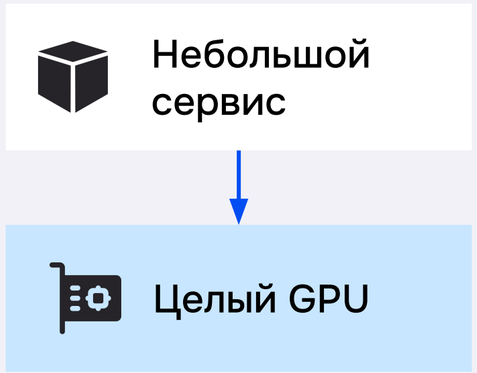
Как выделить GPU для сервиса модели

Сначала оценим память, затем выберем способ выделения

Оборудование куплено, но новые задачи всё равно ждут

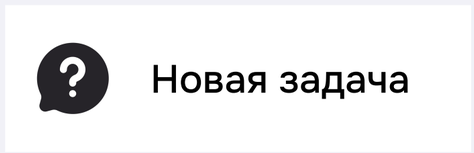
deckhouse kept

Один сервис — целый GPU



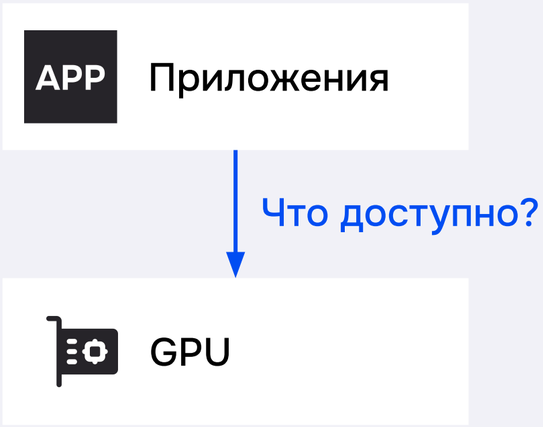
Этот GPU недоступен другим задачам

Нужна настройка узла



Команда ждёт инженера

Не видно доступного ресурса



Трудно обосновать следующую закупку

Память GPU под модель и нагрузку

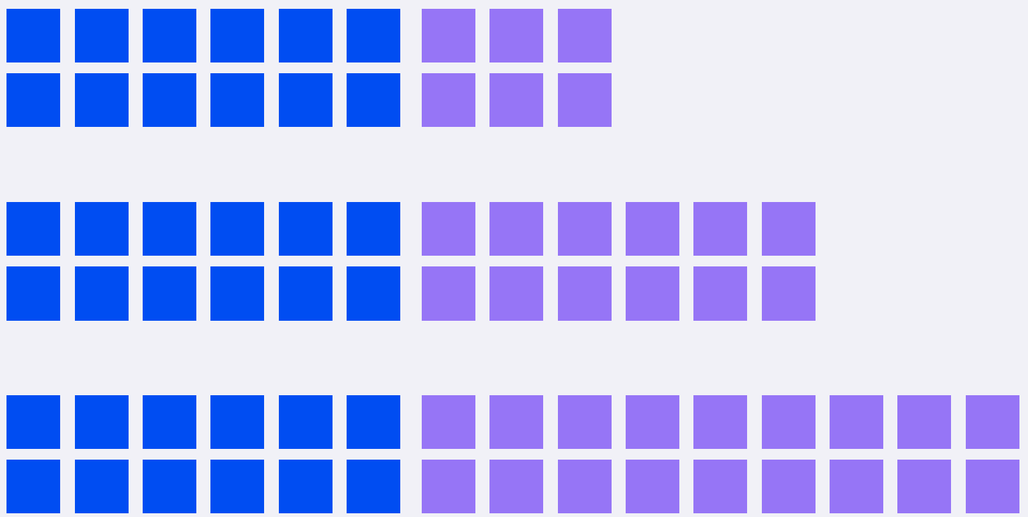
Потребность в памяти оцениваем до запуска сервиса

Языковая модель и формат весов не меняются

Короткий контекст, один запрос

Длинный контекст, один запрос

Длинный контекст, одновременные запросы



В каждом случае:

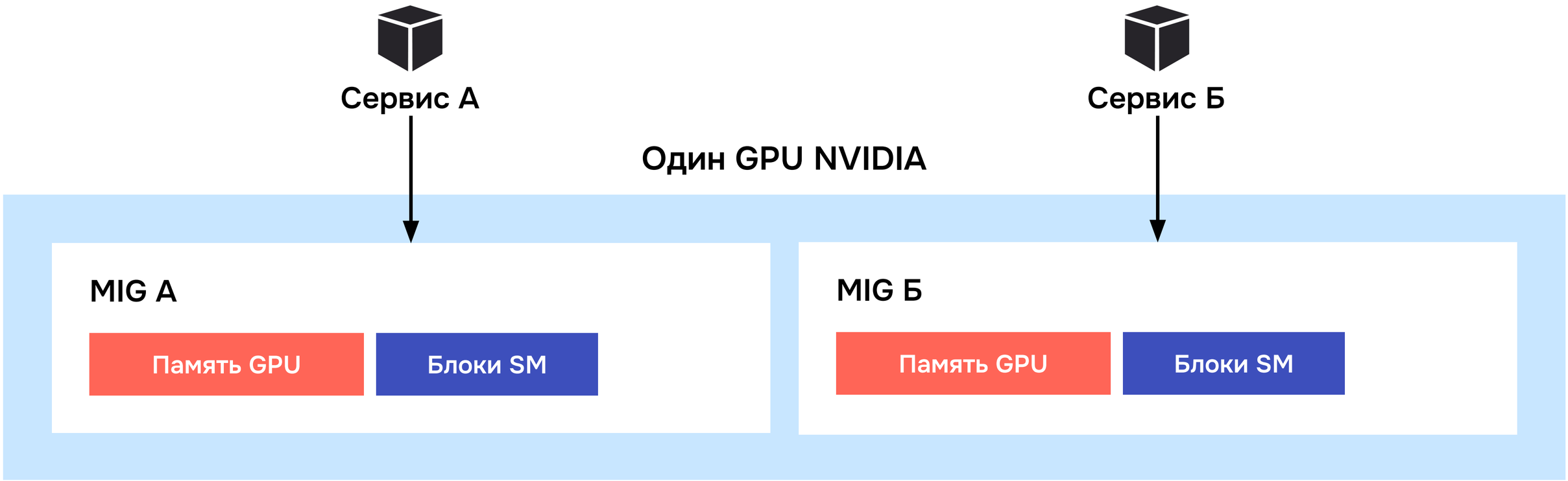
буферы и резерв движка

Веса

KV-кеш

MIG разделяет память и вычислительные блоки

deckhouse kept

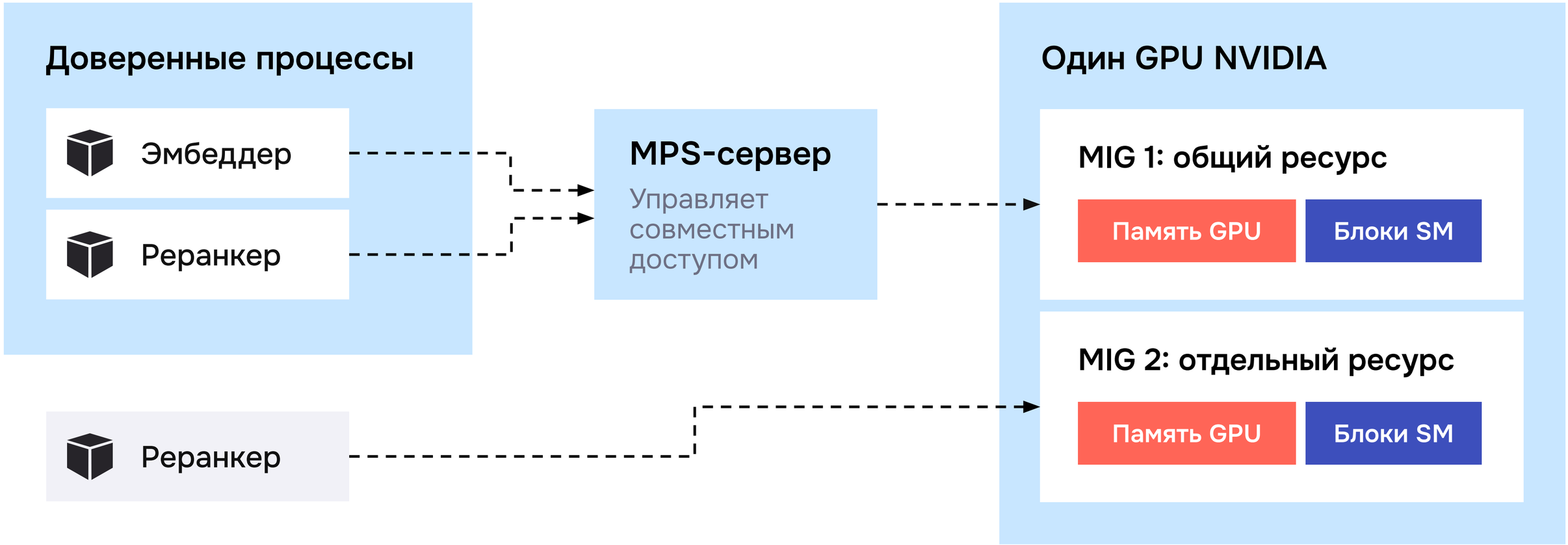


Размеры — из поддерживаемых профилей карты

MPS: несколько процессов используют GPU или MIG

deckhouse kept

Пример: сервисы поиска



Администратор задаёт пул, команда выбирает профиль

deckhouse kept

Администратор

Пул GPU

- Состав пула
- Режимы работы
- Условия совместного использования

Администратор задаёт пул, команда выбирает профиль

deckhouse kept

Администратор

Пул GPU

- Состав пула
- Режимы работы
- Условия совместного использования

Команда

01 Целый GPU

02 Inference-small
Отдельный MIG

03 MIG + MPS

Заявка приложения
Inference-small, 1 MIG

Администратор задаёт пул, команда выбирает профиль

deckhouse kept

Администратор

Пул GPU

- Состав пула
- Режимы работы
- Условия совместного использования

Команда

01 Целый GPU

02 Inference-small
Отдельный MIG

03 MIG + MPS

Заявка приложения
Inference-small, 1 MIG

Платформа

Готовит ресурс
и подключает его

APP Контейнер приложения

✓ MIG подключен

GPU и правила пула

DRA — стандарт Kubernetes для выбора устройств

deckhouse kept

Пример выбора целого GPU

ResourceSlice

Разные GPU в кластере

gpu-01	L40S	48 ГБ
gpu-02	H100 SXM	80 ГБ
gpu-03	B300	288 ГБ

Статья о DRA на Хабре

DRA — стандарт Kubernetes для выбора устройств

deckhouse kept

Пример выбора целого GPU

ResourceSlice

Разные GPU в кластере

gpu-01	L40S	48 ГБ
gpu-02	H100 SXM	80 ГБ
gpu-03	B300	288 ГБ

DeviceClass

Какие устройства подходят

GPU NVIDIA
Целые устройства

Статья о DRA на Хабре

DRA — стандарт Kubernetes для выбора устройств

deckhouse kept

Пример выбора целого GPU

ResourceSlice

Разные GPU в кластере

gpu-01	L40S	48 ГБ
gpu-02	H100 SXM	80 ГБ
gpu-03	B300	288 ГБ

DeviceClass

Какие устройства подходят

GPU NVIDIA
Целые устройства

ResourceClaim

Что нужно приложению

1 целый GPU
Память ≥ 96 ГБ

Статья о DRA на Хабре



DRA — стандарт Kubernetes для выбора устройств

deckhouse kept

Пример выбора целого GPU

ResourceSlice

Разные GPU в кластере

gpu-01	L40S	48 ГБ
gpu-02	H100 SXM	80 ГБ
gpu-03	B300	288 ГБ

DeviceClass

Какие устройства подходят

GPU NVIDIA
Целые устройства

ResourceClaim

Что нужно приложению

1 целый GPU
Память ≥ 96 ГБ

Статья о DRA на Хабре



Планировщик Kubernetes

Выбирает подходящее устройство

Драйвер готовит устройство; среда запуска подключает его

DRA — стандарт Kubernetes для выбора устройств

deckhouse kept

Пример выбора целого GPU

ResourceSlice

Разные GPU в кластере

gpu-01	L40S	48 ГБ
gpu-02	H100 SXM	80 ГБ
gpu-03	B300	288 ГБ

DeviceClass

Какие устройства подходят

GPU NVIDIA
Целые устройства

ResourceClaim

Что нужно приложению

1 целый GPU
Память ≥ 96 ГБ

Статья о DRA на Хабре



Планировщик Kubernetes

Выбирает подходящее устройство

Результат в ResourceClaim

Gpu-03 B300 288 ГБ

Драйвер готовит устройство; среда запуска подключает его

От модели к API: планируемый запуск

deckhouse kept

01

Выбрать модель и версию

Модуль модели:

- Веса модели
- Движок инференса

Одна версия
для всего модуля

От модели к API: планируемый запуск

deckhouse kept

01

Выбрать модель и версию

Модуль модели:

- Веса модели
- Движок инференса

Одна версия
для всего модуля

02

Выбрать GPU и параметры

- Профиль GPU
- Для LLM
- Длина контекста
- Одновременные запросы

От модели к API: планируемый запуск

deckhouse kept

01

Выбрать модель и версию

Модуль модели:

- Веса модели
- Движок инференса

Одна версия для всего модуля

02

Выбрать GPU и параметры

- Профиль GPU
- Для LLM
- Длина контекста
- Одновременные запросы

03

Запустить сервис

Сервис инференса:

- GPU подготовлен
- Модель загружена

От модели к API: планируемый запуск

deckhouse kept

01

Выбрать модель и версию

Модуль модели:

- Веса модели
- Движок инференса

Одна версия для всего модуля

02

Выбрать GPU и параметры

- Профиль GPU
- Для LLM
- Длина контекста
- Одновременные запросы

03

Запустить сервис

Сервис инференса:

- GPU подготовлен
- Модель загружена

04

Подключить приложение

Указать адрес API и ключ доступа

Кому доступна модель и кто создаёт нагрузку

deckhouse kept

Стенд Deckhouse

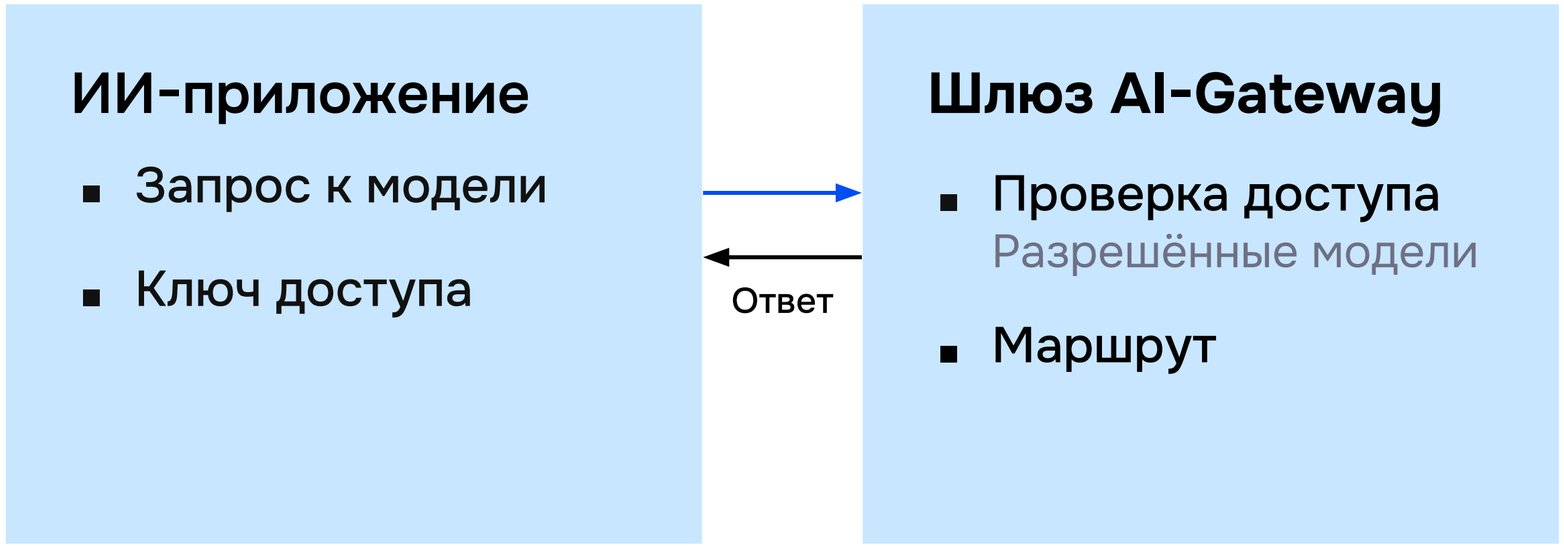
ИИ-приложение

- Запрос к модели
- Ключ доступа

Кому доступна модель и кто создаёт нагрузку

deckhouse kept

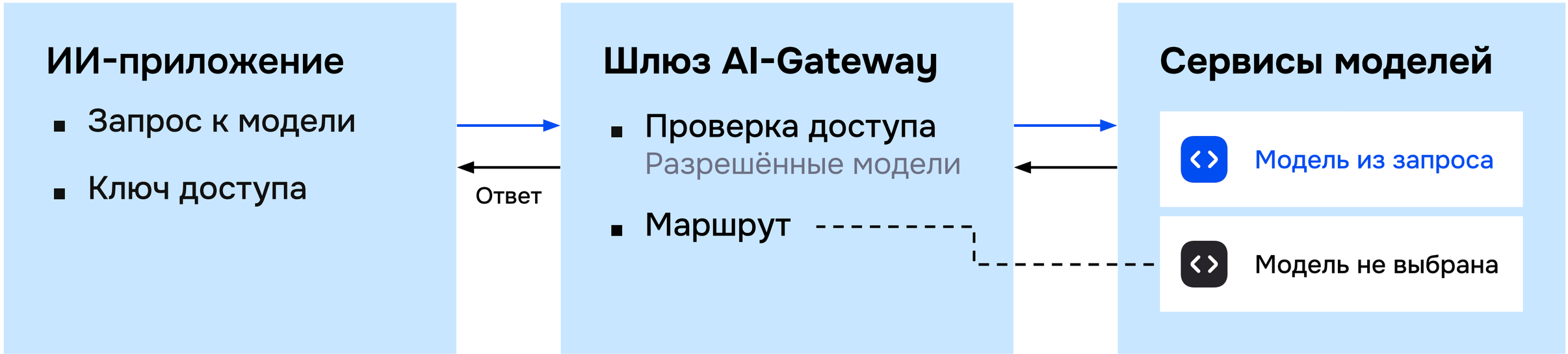
Стенд Deckhouse



Кому доступна модель и кто создаёт нагрузку

deckhouse kept

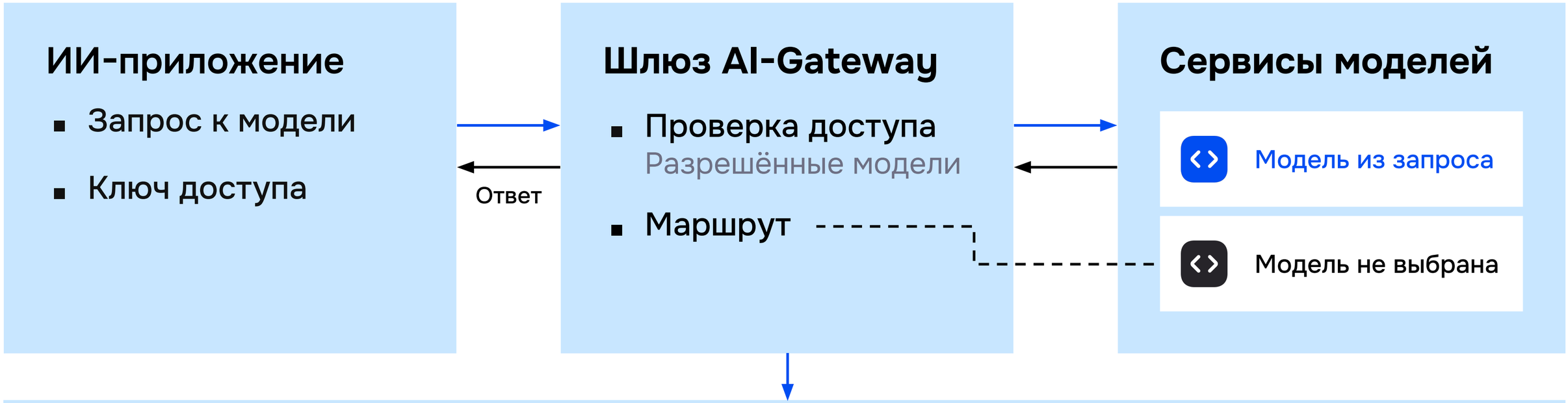
Стенд Deckhouse



Кому доступна модель и кто создаёт нагрузку

deckhouse kept

Стенд Deckhouse



Журнал запросов

Условная запись

Ключ приложения	Модель	Запрос/ответ
app-docs	llm-docs	1200/180 токенов

Сопоставляем метрики сервиса и GPU

deckhouse kept

Сравниваем за один период

Сервис модели

llm-docs

Время ответа

Запросы за период

Ошибки запросов

Сопоставляем метрики сервиса и GPU

deckhouse kept

Сравниваем за один период

Сервис модели

llm-docs

Время ответа

Запросы за период

Ошибки запросов

ResourceClaim

llm-docs-gpu

Выделенный GPU

gpu-03

Сопоставляем метрики сервиса и GPU

deckhouse kept

Сравниваем за один период

При совместном использовании метрики всей карты — общие

Сервис модели

llm-docs

- Время ответа
- Запросы за период
- Ошибки запросов

ResourceClaim

llm-docs-gpu

Выделенный GPU

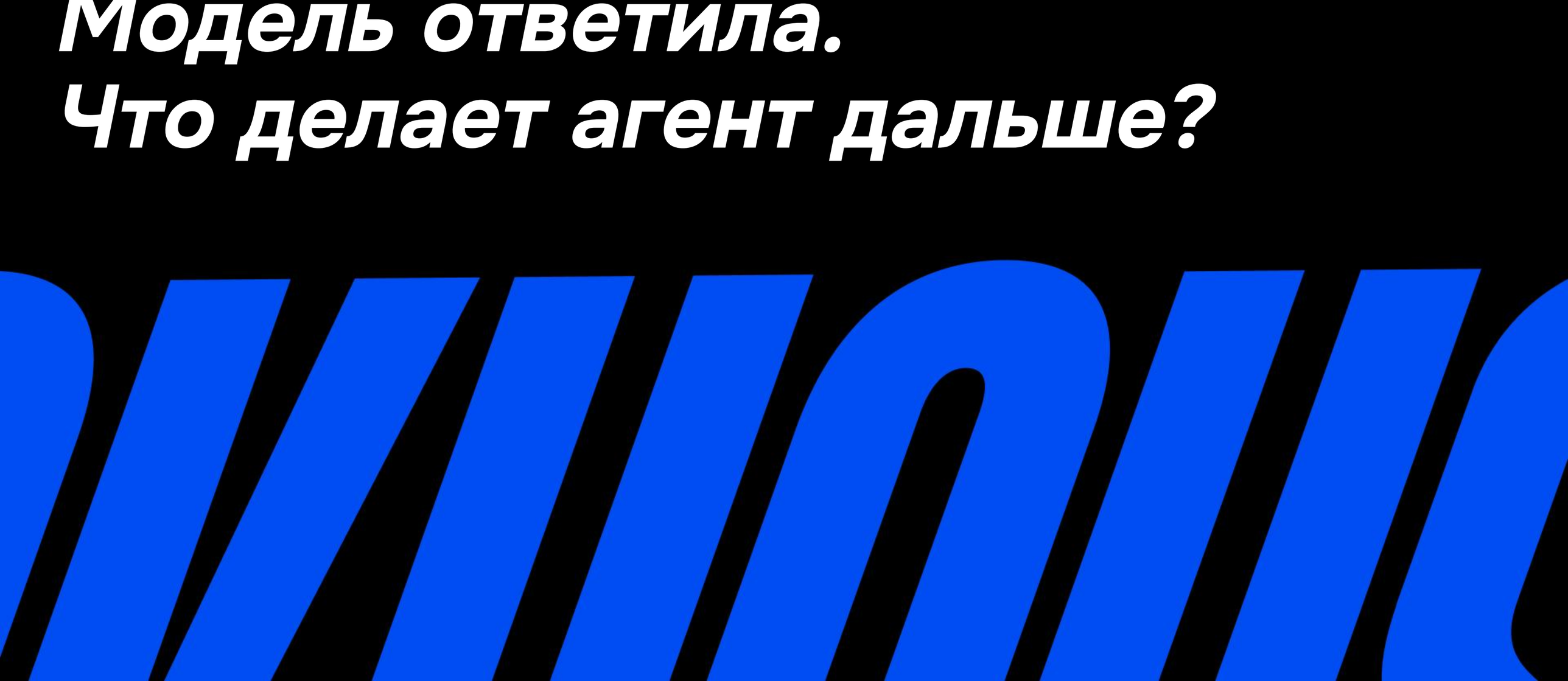
gpu-03

- Вычисления
- Память
- Состояние

Kept: выполнение задания в 1С

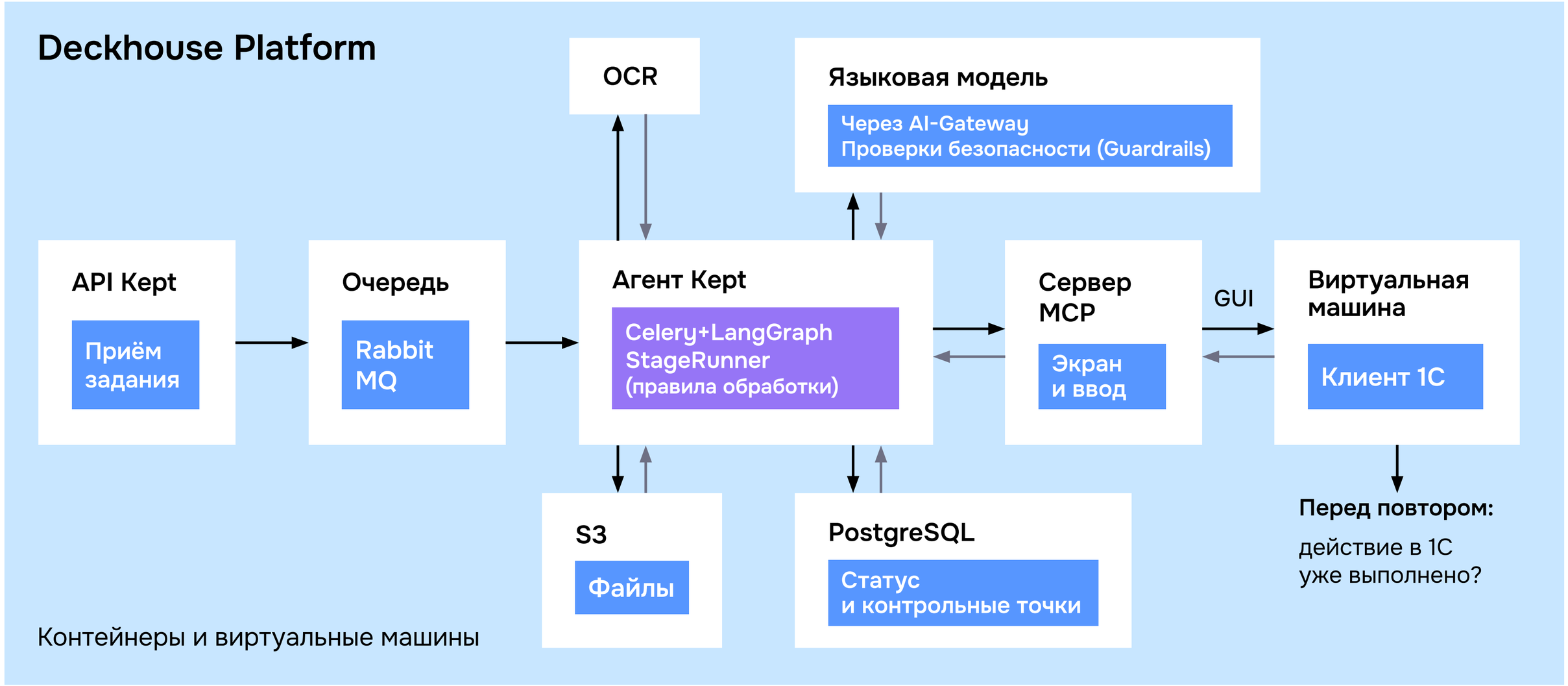
deckhouse kept

Модель ответила.
Что делает агент дальше?



Как агент работает с моделями и 1С

deckhouse kept



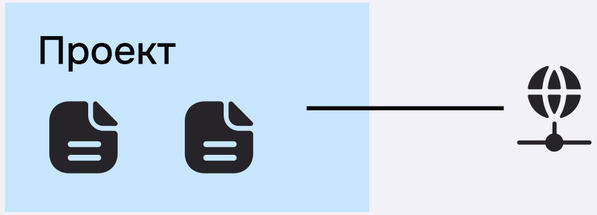
Кто задаёт права агента

deckhouse kept

Проверяем доступ к секретам
и журнал действий агента

Deckhouse Platform

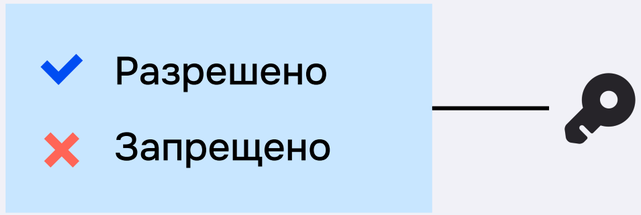
Администратор платформы



- Ресурсы проекта
- Сетевой доступ

Инструменты Kept

Команда приложения



- Разрешённые операции
- Учётные данные

Учётная система

Владелец системы



- Доступ к данным
- Право на изменение

Где искать причину задержки

deckhouse kept

Измеряем весь путь: от документа до завершения операции в 1С

Сервис ещё не готов

- Выдача GPU
- Готовность сервиса и рабочего места

API доступен, ответ запаздывает

- Очередь запросов
- Вычисления модели

Модель ответила, задача не закончена

- Отклик 1С и действия в GUI
- Сверка и исправления

Выбираем процессы для автоматизации

deckhouse kept

Процесс	Потенциал, штатные единицы	Сложность, балл
Учёт приобретений: активы, запасы, работы и услуги	30	4
Авансовые отчёты	8	4
Реализация готовой продукции и прочих активов	7	4
Перемещение запасов	12	5
Оформление приёма на работу	6	3
Списание запасов	4	3

Всего — 67 штатных единиц

Сколько ручной работы можно автоматизировать

deckhouse kept

Учёт приобретений: 1,46 млн пакетов документов в год

Подразделения в проекте

88

штатных единиц

Всего в процессе:
90 штатных единиц

×

Документы для автоматизации

49 %

всех документов

Автоматизируем:
70 % по запасам
50 % по услугам

×

Учётные системы в проекте

69 %

всех документов

≈1 млн из 1,46 млн пакетов документов

≈

Потенциал высвобождения

30

штатных единиц

Затраты и экономия за пять лет

deckhouse kept

Финансовая модель четырёх процессов на 2027–2031

Затраты на проект

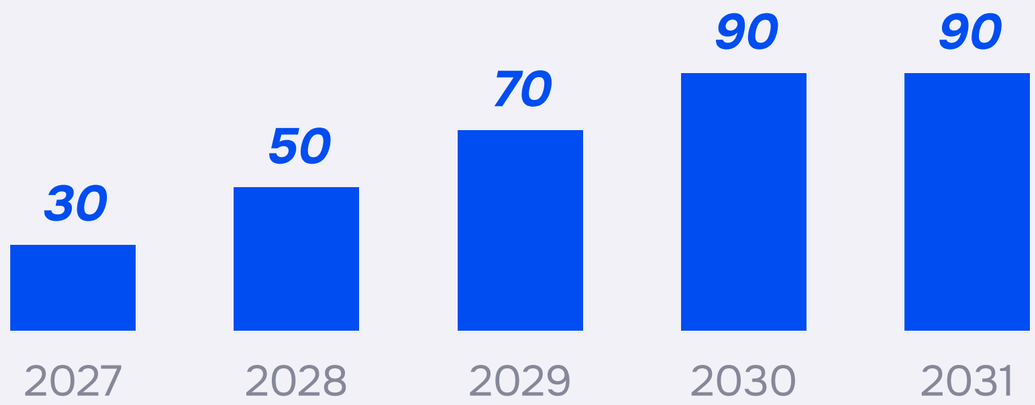
200 млн ₽

Консалтинг	**90 млн ₽**
Инфраструктура	**60 млн ₽**
Внутренняя команда	**50 млн ₽**

Плановая экономия

825 млн ₽

Затраты на штатную единицу — 2,5 млн ₽ в год

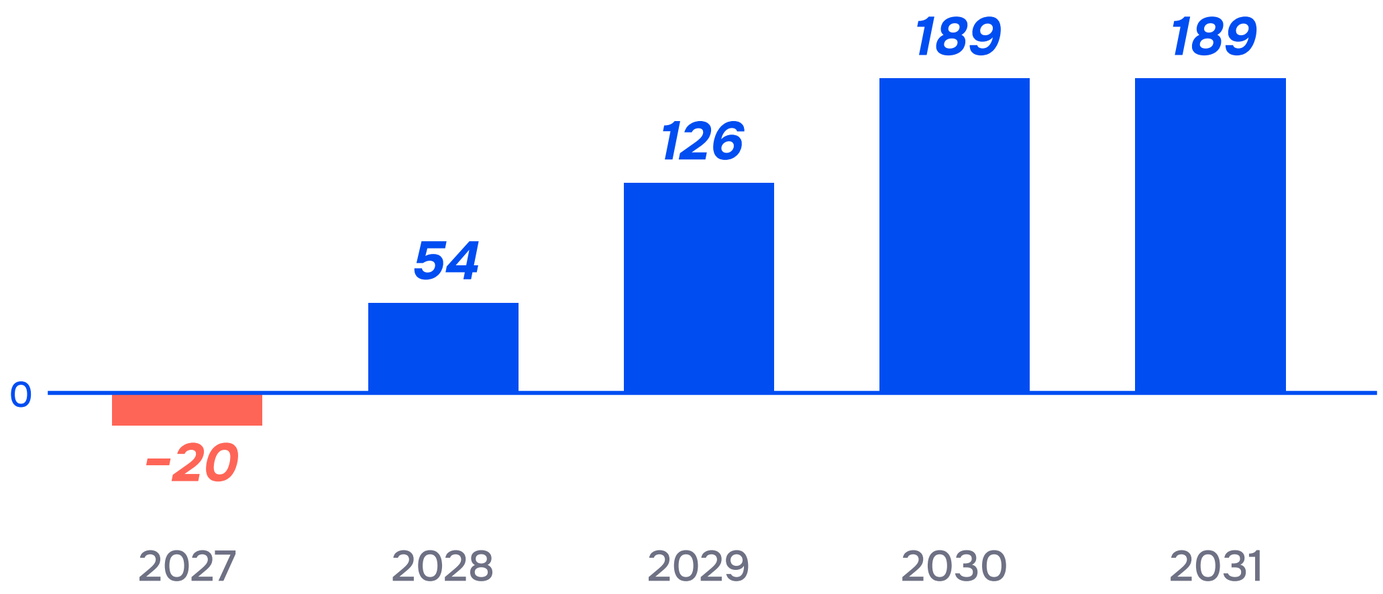


Денежный поток и окупаемость

deckhouse kept

Прогноз денежного потока, млн ₽

После затрат и налогов, с поправкой на инфляцию 10 %



Чистый денежный поток за пять лет

538 млн ₽

Чистая приведённая стоимость (NPV)	234 млн ₽
Внутренняя норма доходности (IRR)	80,9 %
Срок окупаемости	≈1,4 года

Выводы

deckhouse kept

01 Измеримый эффект инициативы

02 Сверхдетальная аналитика процесса и постановка + объёмы тестовых данных = точная оценка GPU-ресурсов и безопасность

03 Платформа управляемых микросервисов на K8s = «офис» ИИ-сотрудников

Разберём ваш агентский проект

deckhouse kept

- Определим, что можно автоматизировать
- Проверим решение на стенде
- Оценим эффект для бизнеса

aleksandr.podmoskovniy@flant.com

RMalyuga@kept.ru

Deckhouse Platform

KeptStore