Лекция ВШЭ · 23 мая 2026

# AIOps / LLMOps для ИИ-продуктов

Управленческие решения, экономика и FinOps для продакшен-контура ИИ

# Александр Поломодов

Спикер Контекст

Technical Director & Fellow, T-Технологии

- Архитектура и инженерные практики
- Внедрение AI на масштабе разработки
- Фокус: управляемый продакшен AI

Вступление | Продолжение

# Платформа есть. Теперь ей надо управлять

Руководитель выбирает поведение, запуск и контроль

**Словарь уже есть**

Шлюз, контекст, evals и защитные ограничения известны.

**Фокус на решениях**

SLA, автономия и цена качества.

**Денежный якорь**

Бот поддержки проверяет TCO/FinOps.

Навигация Что не дублируем

# Базовые компоненты считаем пройденными

AIOps / LLMOps начинается после появления платформы

- Не пересказываем карту платформы
- Выбираем сервис под сценарий
- Считаем цену решенной задачи

- Фокус: управляемость AI-продукта



Блок 1

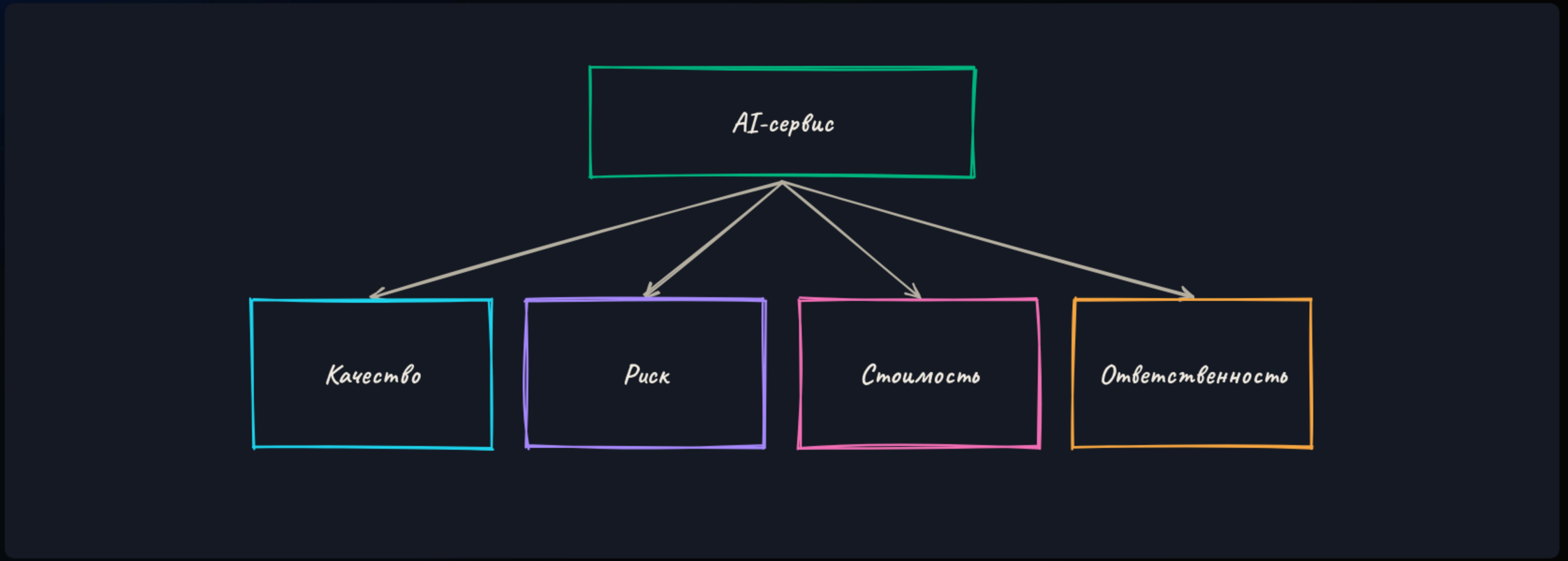
01

# От базы к управлению

Не повторяем устройство AI-платформы, а переводим его в решения руководителя

Обзор

# После запуска AI становится сервисом



# Менеджер выбирает рамку



## Цель

Сценарий
Базовая линия

## Границы

Цена ошибки
Стоимость операции

# AIOps — это контур решений







Блок 2

02

# Решения менеджера

Сценарий, автономия, цена ошибки, SLA, бюджет качества и ответственность

# До масштаба нужны решения

Иначе AI растет быстрее контроля

Ценность и сценарий

## Ценность

Где ценность, базовая линия и SLA.

Риск и контроль

## Риск

Допустимая автономия и участие человека.

Экономика и ответственность

## Стоимость

Цена качества; владельцы продукта, данных, риска и счета.

# Цена ошибки задает автономию

| Низкий риск | Высокий риск |
| --- | --- |
| Черновик · поиск | Деньги · продакшен |
| Быстрая модель | Согласование + аудит |

# AI SLA шире uptime API



## Сервис

Задержка и тайм-аут

Очереди и резервный сценарий

## Качество

Источники и полнота

Политики и эскалация

# Оценка идет от продукта

Одинаковый стек может быть правильным или нет

- Ценность: процесс и базовая линия
- Риск: ошибка и согласование
- Экономика: цена единицы
- Операционность: выпуск и откат

- Решение объясняется не стеком



Блок 3

03

# Варианты и компромиссы

API-first, управляемая платформа, арендованные GPU, свой self-host и гибрид

# AI-runtime — портфель вариантов

| API / SaaS | Управляемая платформа | Аренда GPU | Свои GPU |
|---|---|---|---|
| Скорость<br>Готовый сервис | Контроль<br>и зависимость<br>от поставщика | Своя модель<br>OPEX | Контроль<br>CAPEX +<br>эксплуатация |

Гибрид распределяет разные сценарии по разным путям

# Помощник и агент — разные продукты

## Помощник

Читает и объясняет
ACL + источники

## Агент

Выполняет действия
Права + согласование

Маршрутизация

# Нужны правила маршрутизации



Блок 3 · Вывод Компромисс

# Выбор — портфельное решение

- API-first — старт и проверка
- Аренда GPU — переходный контроль
- Свой self-host — только при нагрузке
- Гибрид — разные сценарии

- Нельзя выбрать рантайм без сценариев и финансового горизонта



Блок 4

04

# Денежный кейс

Бот поддержки: нагрузка, токены, GPU, команда и управленческий выбор

Бот поддержки

# Считаем support-бота

200 000 диалогов в месяц, 10 сообщений в диалоге, SLA до 5 секунд

**Что считаем**

Нагрузка, токены, пик, GPU, CAPEX/OPEX, команда.

**Что сравниваем**

API-first, аренда GPU, свой self-host, гибрид.

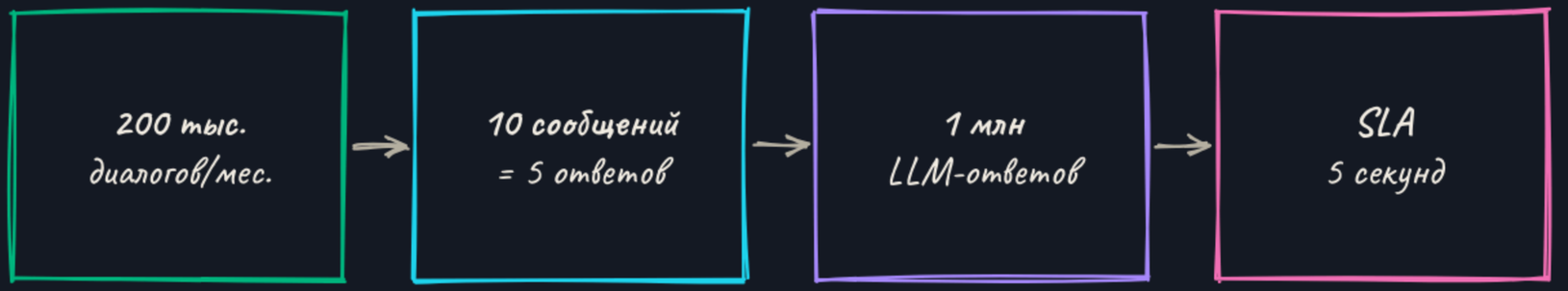
**Что решаем**

Путь на год и метрики пересмотра.

# Параметры кейса фиксируют масштаб решения

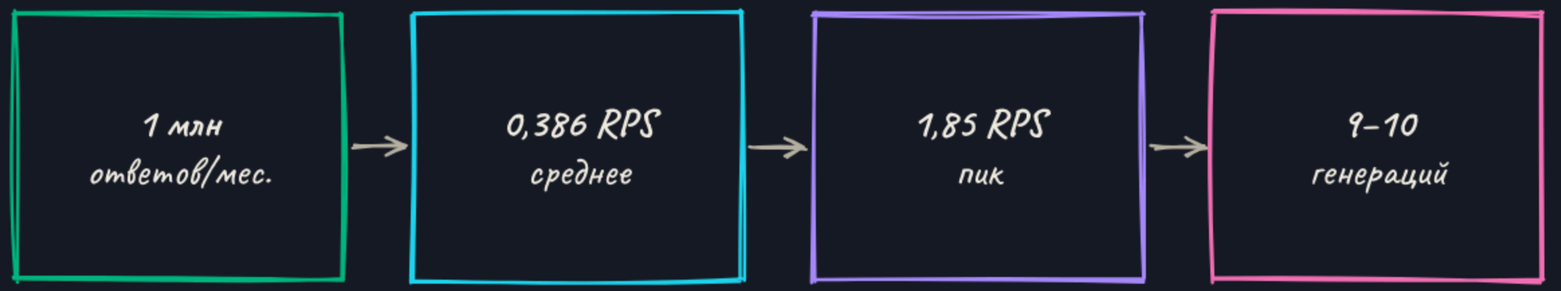
Кейс · Входные параметры

200 000 диалогов



# Переводим параметры в RPS и параллельность

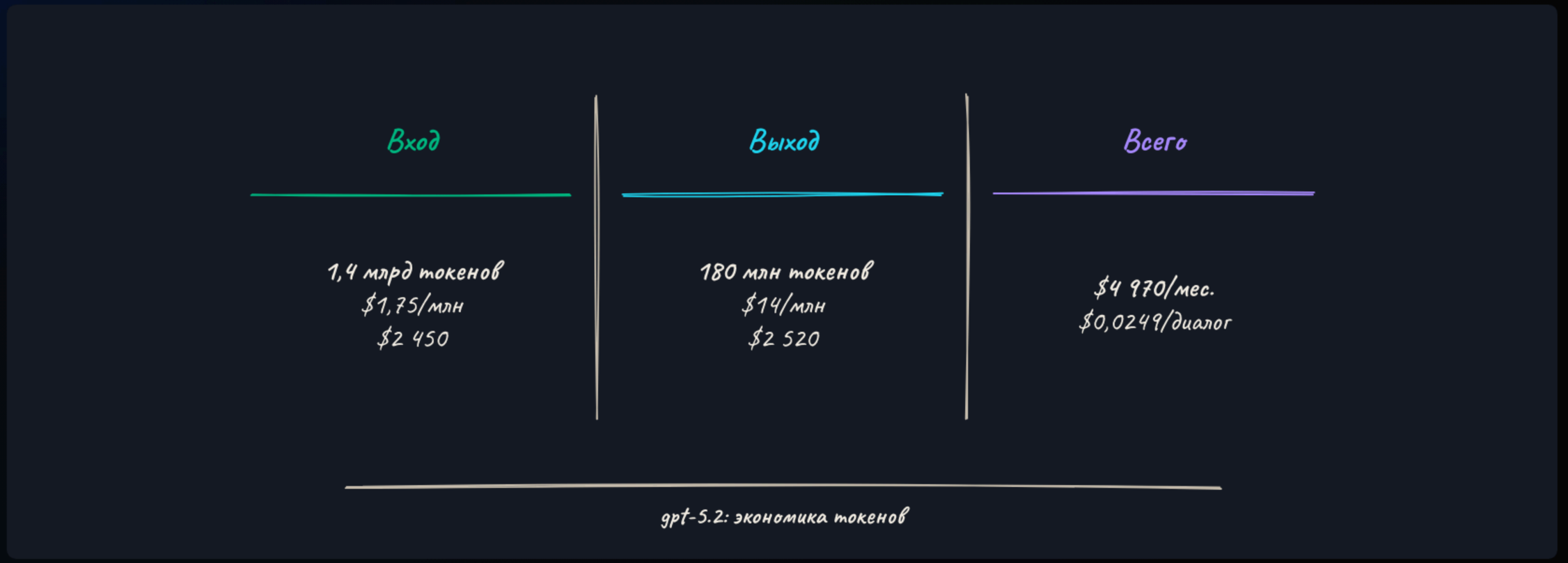
Кейс · Шаг 1 Нагрузка




Параллельность ≈ пиковый RPS × 5 секунд

# GPT-5.2 API дает понятный стартовый OPEX

Кейс · Шаг 2 API-first



# API-first покупает скорость

Кейс · API-first

Взгляд менеджера

Риск — тихий рост юнит-экономики

+ Почему стартовать удобно

- Почти нулевой CAPEX
- Быстрее в продакшен
- Тоньше платформа вокруг

! Что нужно контролировать

- Выходные токены, кеш, повторы
- Лимиты провайдера и деградация

# Self-host считают от SLA и пиков



| Qwen-14B · INT4 | Qwen-32B · INT4 |
| --- | --- |
| ~90 ток/с | ~45 ток/с |
| 180 ток: 2,0 с | 180 ток: 4,0 с |
| ~2 запроса/GPU | ~1 запрос/GPU |
| 6–8 GPU | ≈12 GPU |

Пик ≈10 ответов, SLA 5 с; с батчингом GPU понадобится меньше

# Аренда GPU меняет CAPEX на OPEX




Ставка — по прайс-листу cloud.ru, февраль 2026; сеть и резерв сверху

# Self-host меняет задачу



| Аренда | Свои GPU |
| --- | --- |
| Проверка гипотезы | Стабильная нагрузка |
| OPEX и MLOps | CAPEX и поставка |

# Сравнение стратегий: решение на первый год



| | API | Аренда GPU | Свои GPU |
|---|---|---|---|
| CAPEX | Низкий | Низкий | Высокий |
| OPEX | ~$5 тыс. LLM/мес. | 1,17–2,33 млн ₽/мес. | Инфраструктура + команда |
| Запуск | Быстро | Средне/быстро | Медленнее |
| Контроль | Ниже | Высокий | Максимум |

При текущей нагрузке — API-first

# Решение должно включать точку пересмотра

Кейс · Решение 12 месяцев




Обязательное размещение на своих серверах меняет выбор

# Решение по кейсу

Кейс · Вывод Ответ менеджера

- Стартовать API-first
- Сразу вести юнит-экономику
- Задать триггеры пересмотра
- Self-host доказывать расчетом

- Первое решение - скорость с контролем, не вечная архитектура



Блок 5

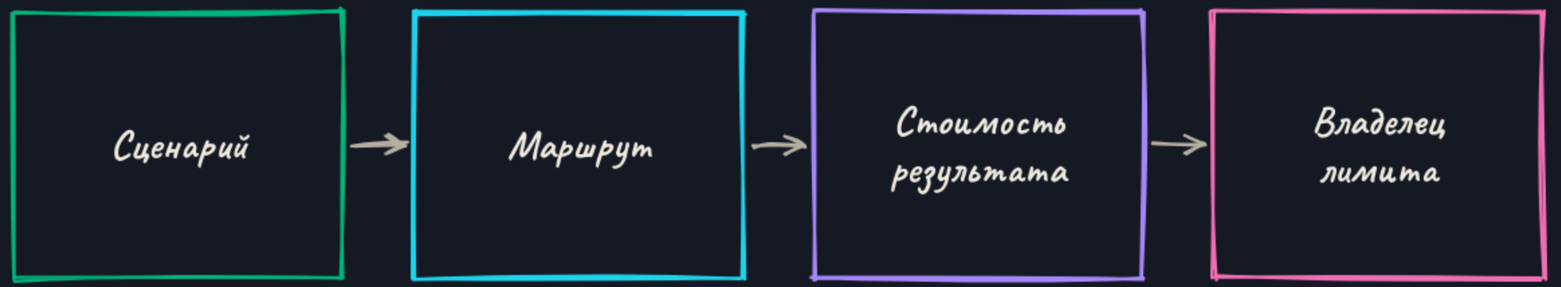
05

# Управляемость и FinOps

Дашборд, распределение затрат, пересчет юнит-экономики и 120-дневный план

Управление

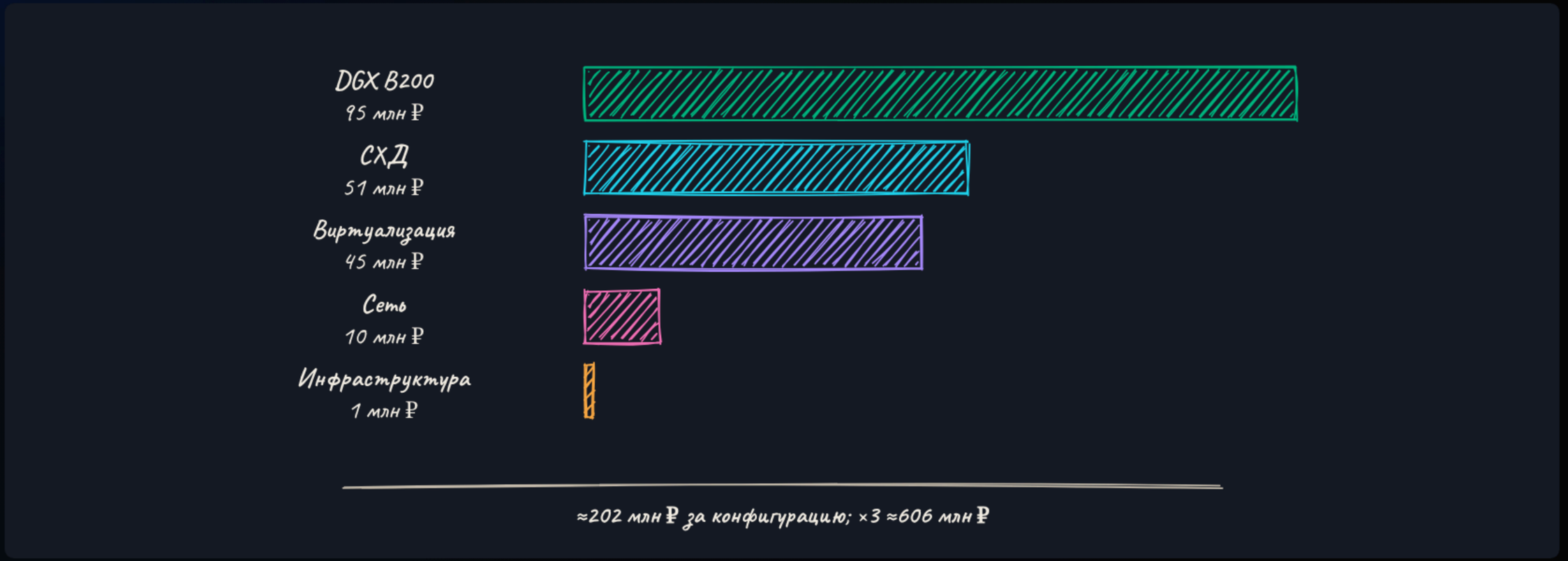
# FinOps начинается с владельца счета



FinOps · Сверка с реальностью

CAPEX

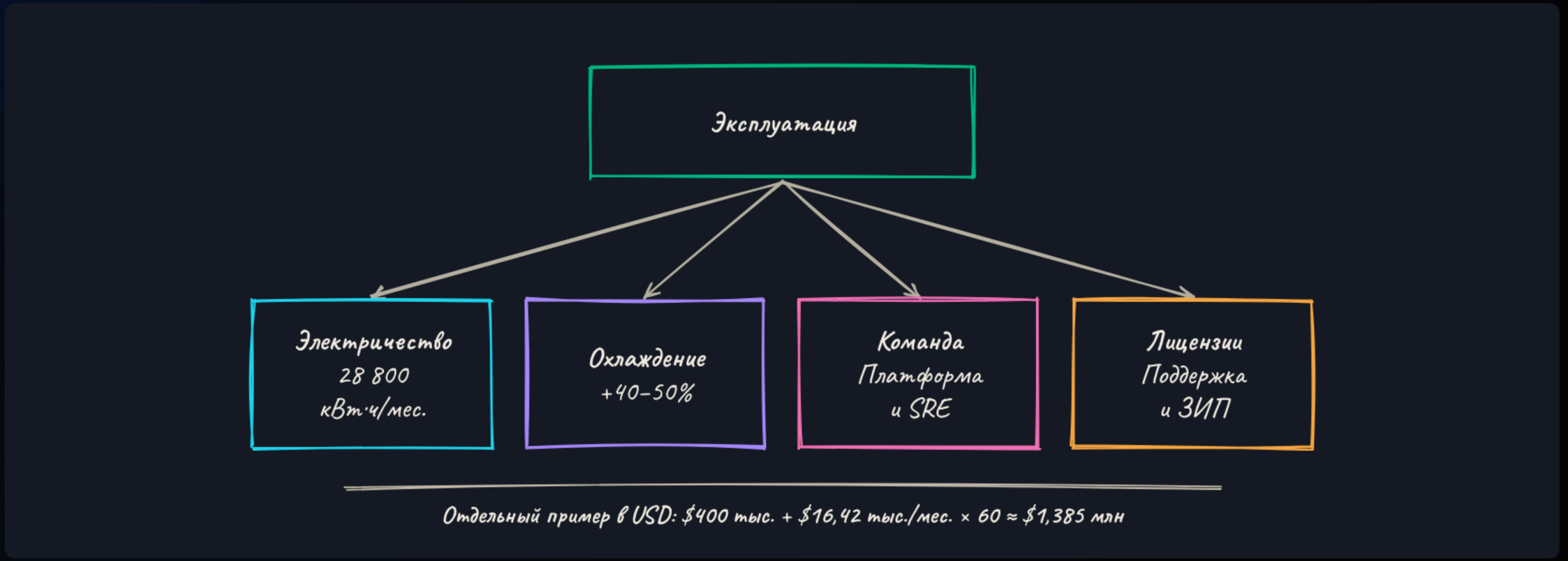
# Своё железо превращает счёт в платформенный



# После CAPEX начинается эксплуатация

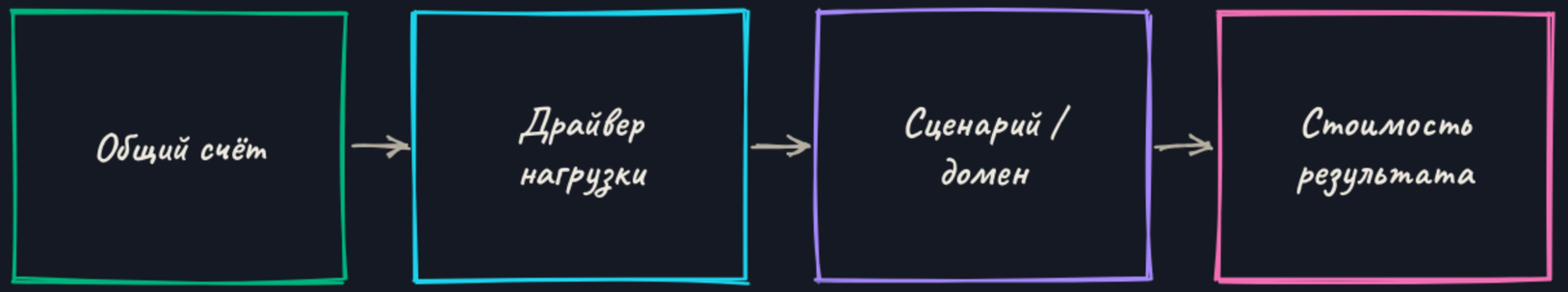
FinOps · OPEX

После покупки



Кому счёт

# Общую платформу делят по понятным драйверам



Блок 5 · Дашборд

Что видит менеджер

# Дашборд соединяет качество, риск и деньги

## Качество и риск

Сценарии

Резервный сценарий и политики

Инциденты

## Экономика

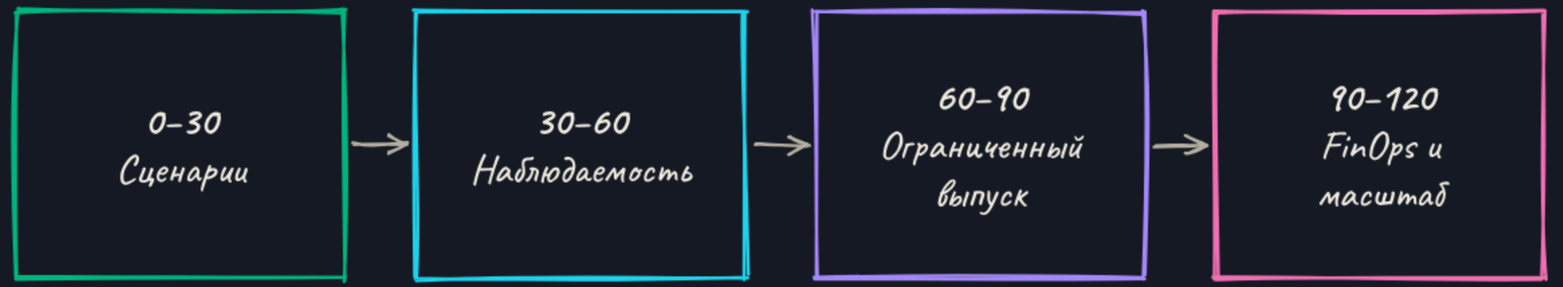
Стоимость результата

Кеш и повторы

Состав маршрутов

# 120 дней: от сценария к масштабу

Финал · Дорожная карта 120 дней



Финал

Главный вопрос

# Нельзя измерить — нельзя управлять

- Без evals качество кажется выше
- Без резервного сценария растет риск
- Без юнит-экономики ломается экономика
- Без ответственности начинается спор

- Управляемый ИИ начинается с решений менеджера

Источники Литература

# Материалы

AIOps, риск и экономика

- OpenAI: цены и документация.
- OWASP, NIST, Langfuse.
- Habr TCO, FinOps Foundation.
- Прайс-лист ↗cloud.ru, бенчмарк Qwen2.5.

ФОКН

# Форма обратной связи по второму дню

