SAINT HIGHLOAD++ 2026

«Давайте просто развернём свою LLM на 1T параметров»

как выглядит self-hosted AI после первого миллиона запросов

Сергей Нотевский, Битрикс24

Saint HighLoad++ 2026

Кто рассказывает

Сергей Нотевский

AI Platform Lead, Битрикс24

Вместе с командой построил AI-платформу Битрикс24 с нуля - до 15 млн+ запросов в месяц на 200+ GPU.

AI в Битрикс24

CRM

- транскрибация звонков
- саммари звонка
- заполнение карточки
- оценка качества разговора
- скоринг лидов

Чаты

- ассистент
- ответы по базе знаний

Сайты

- генерация сайта с нуля
- редактирование блоков сайта

Встречи и звонки

- итоги встречи
- договорённости
- поручения

Задачи

- постановка из чата
- из аудио- и видеосообщений
- поиск просрочек
- анализ потока задач

Агенты

- действия внутри Битрикс24
- конструктор агентов

Технологические партнёры

- работа в досках
- голосовой ассистент
- VibeCode платформа
- внешние агенты (Hermes, OpenClaw)

Внутренние команды

- первая линия поддержки
- ревью кода
- QA-тесты
- создание контента (маркетинг)
- AI-тренажёры

AI в Битрикс24

CRM

- транскрибация звонков
- саммари звонка
- **заполнение карточки**
- оценка качества разговора
- скоринг лидов

Чаты

- ассистент
- ответы по базе знаний

Сайты

- генерация сайта с нуля
- редактирование блоков сайта

Встречи и звонки

- итоги встречи
- договорённости
- поручения

Задачи

- постановка из чата
- из аудио- и видеосообщений
- поиск просрочек
- анализ потока задач

Агенты

- действия внутри Битрикс24
- конструктор агентов

Технологические партнёры

- работа в досках
- голосовой ассистент
- VibeCode платформа
- внешние агенты (Hermes, OpenClaw)

Внутренние команды

- первая линия поддержки
- ревью кода
- QA-тесты
- создание контента (маркетинг)
- AI-тренажёры

AI в Битрикс24

CRM

- транскрибация звонков
- саммари звонка
- заполнение карточки
- оценка качества разговора
- скоринг лидов

Чаты

- ассистент
- ответы по базе знаний

Сайты

- **генерация сайта с нуля**
- редактирование блоков сайта

Встречи и звонки

- итоги встречи
- договорённости
- поручения

Задачи

- постановка из чата
- из аудио- и видеосообщений
- поиск просрочек
- анализ потока задач

Агенты

- действия внутри Битрикс24
- конструктор агентов

Технологические партнёры

- работа в досках
- голосовой ассистент
- VibeCode платформа
- внешние агенты (Hermes, OpenClaw)

Внутренние команды

- первая линия поддержки
- ревью кода
- QA-тесты
- создание контента (маркетинг)
- AI-тренажёры

AI в Битрикс24

CRM

- транскрибация звонков
- саммари звонка
- заполнение карточки
- оценка качества разговора
- скоринг лидов

Чаты

- ассистент
- ответы по базе знаний

Сайты

- генерация сайта с нуля
- редактирование блоков сайта

Встречи и звонки

- итоги встречи
- договорённости
- поручения

Задачи

- постановка из чата
- из аудио- и видеосообщений
- поиск просрочек
- анализ потока задач

Агенты

- **действия внутри Битрикс24**
- конструктор агентов

Технологические партнёры

- работа в досках
- голосовой ассистент
- VibeCode платформа
- внешние агенты (Hermes, OpenClaw)

Внутренние команды

- первая линия поддержки
- ревью кода
- QA-тесты
- создание контента (маркетинг)
- AI-тренажёры

Платформа в цифрах

15M+

LLM-запросов в месяц

2с - 2мин

время жизни запроса

4,5 млрд

токенов в сутки

200+

GPU четырёх поколений

20+

команд-потребителей

Почему мы ушли с MaaS?

Зачем все эти заморочки?

Почему мы ушли с MaaS

Данные

Обработка - только в своём техконтуре: это обещание клиентам продукта.

Почему мы ушли с MaaS

Данные

Обработка - только в своём техконтуре: это обещание клиентам продукта.

Гибкость

Свои модели, лимиты и приоритеты под реальные сценарии.

Почему мы ушли с MaaS

Данные

Обработка - только в своём техконтуре: это обещание клиентам продукта.

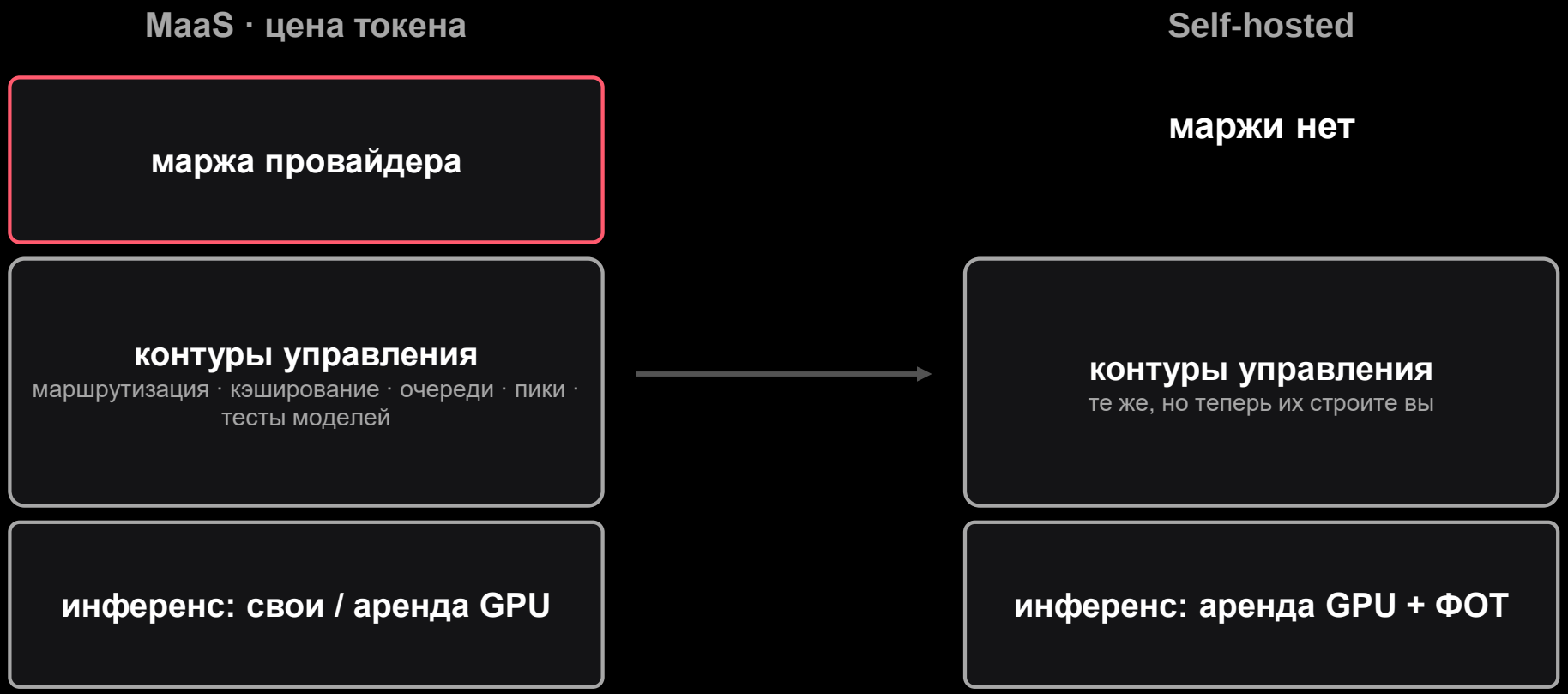
Гибкость

Свои модели, лимиты и приоритеты под реальные сценарии.

Экономика

На большом стабильном объёме токенов своё железо окупается.

Из чего состоит цена токена MaaS



В MaaS платите за маржу провайдера и его контуры. В self-hosted маржи нет, но контуры и железо (аренда + ФОТ) теперь ваши.

О чём этот доклад?

О чём этот доклад?

Пять кейсов, в которых мы наступили на грабли и серьёзно перестроили платформу.

КЕЙС 1 ИЗ 5

Карточка модели не равна модели в продакшене

«Она поддерживает миллион токенов контекста и влазит в H100. Просто разверните»

gpt-oss-120b: карточка против продакшена

В карточке

- контекст до 1M токенов
- запускается на 1×H100
- создана для агентных задач

gpt-oss-120b: карточка против продакшена

В карточке

- контекст до 1M токенов
- запускается на 1×H100
- создана для агентных задач

В проде

- ~120K контекста - потолок памяти на H100
- после 64K качество резко падает
- Responses API не закрыл наши agent flows

Релиз весов ≠ готовность к продакшену

День 0

Веса вышли

пост в блоге: «vLLM уже поддерживает»

Релиз весов ≠ готовность к продакшену

День 0

Веса вышли

пост в блоге: «vLLM уже поддерживает»

Недели 1-2

Первые грабли

MXFP4 (родной формат квантования) - только на Hopper+; нужного API инференса нет

Релиз весов ≠ готовность к продакшену

День 0

Веса вышли

пост в блоге: «vLLM уже поддерживает»

Недели 1-2

Первые грабли

MXFP4 (родной формат квантования) - только на Hopper+; нужного API инференса нет

Недели 2-4

Патчим сами

патч vLLM, обход через Triton backend

Релиз весов ≠ готовность к продакшену

День 0

Веса вышли

пост в блоге: «vLLM уже поддерживает»

Недели 1-2

Первые грабли

MXFP4 (родной формат квантования) - только на Hopper+; нужного API инференса нет

Недели 2-4

Патчим сами

патч vLLM, обход через Triton backend

~Месяц

Релиз в продукт

проверка качества → пробный трафик → алиас модели

Кейс GPT-OSS

Родное квантование MXFP4 заводилось только на Hopper+, а пул под модель у нас на A100. Responses API в vLLM не было. Сделали обход через Triton и несли патч vLLM, пока поддержка не доехала в основной vLLM.

Это случалось не раз

GPT-OSS 120B

на A100 не было нужного API инференса - обход и свои патчи vLLM

Это случалось не раз

GPT-OSS 120B

на A100 не было нужного API инференса - обход и свои патчи vLLM

Qwen3-Next

prefix cache в vLLM ждали 3 месяца - 5 минут на «привет, как дела?»

Это случалось не раз

GPT-OSS 120B
на A100 не было нужного API инференса - обход и свои патчи vLLM

Qwen3-Next
prefix cache в vLLM ждали 3 месяца - 5 минут на «привет, как дела?»

Kimi 2.6
свой формат вызова инструментов и строгий JSON-ответ - снова патчи в форке

А бизнес ждёт поставки модели «ещё вчера!»

Итоговый time to prod - от пары дней до нескольких месяцев.

Гибридная модель с reasoning в проде хуже пары моделей

Один гибрид

instruct и reasoning-режим в одном инстансе

- качество ниже в обоих режимах
- reasoning-трафик давит короткие запросы
- разбор reasoning-разметки тормозит даже обычные ответы
- один пул: лимиты не развести

Гибридная модель с reasoning в проде хуже пары моделей

Один гибрид

instruct и reasoning-режим в одном инстансе

- качество ниже в обоих режимах
- reasoning-трафик давит короткие запросы
- разбор reasoning-разметки тормозит даже обычные ответы
- один пул: лимиты не развести

Пара: Instruct и Reasoning

две модели, два пула, два профиля нагрузки

- качество выше на наших датасетах
- нагрузки изолированы по пулам
- лимиты, масштаб, выкатки независимые

«Ладно, давай развернём одну большую модель на 1Т!»

Она же мощнее! Должна больше обрабатывать!

Сколько стоит своя LLM на 1T параметров

Kimi 2.6

отлично справляется с агентами, но…

Сколько стоит своя LLM на 1T параметров

Kimi 2.6

отлично справляется с агентами, но…

Нужен сервер 8×H200

дороже примерно в 8-10 раз, чем сервер для gpt-oss (1×H100), но…

Сколько стоит своя LLM на 1T параметров

Kimi 2.6
отлично справляется с агентами, но…

Нужен сервер 8×H200
дороже примерно в 8-10 раз, чем сервер для gpt-oss (1×H100), но…

Переваривает в стоке в ~4 раза больше
то есть ×2,5 за единицу capacity, но…

Сколько стоит своя LLM на 1T параметров

Kimi 2.6
отлично справляется с агентами, но…

Нужен сервер 8×H200
дороже примерно в 8-10 раз, чем сервер для gpt-oss (1×H100), но…

Переваривает в стоке в ~4 раза больше
то есть ×2,5 за единицу capacity, но…

Намного сложнее достать
для работы в production нужно минимум 2 инстанса, а ещё dev и test

КЕЙС 2 ИЗ 5

Большие запросы душат маленькие

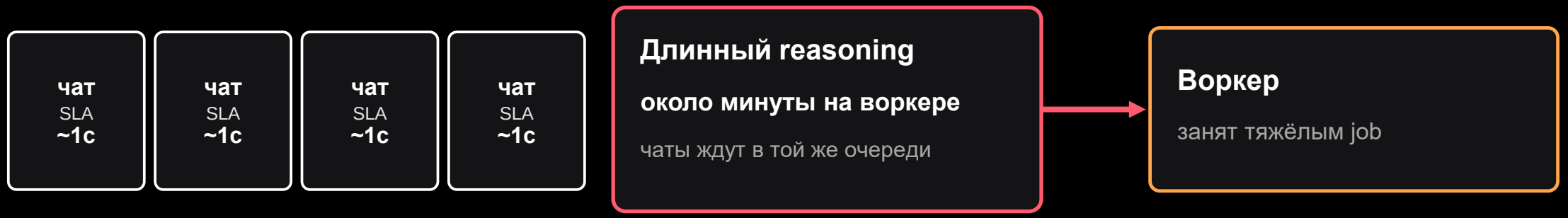
В час пик чаты начали тормозить

На тестах - отлично
нагрузочные прогоны зелёные

Железо работает
GPU не перегружены, ошибок нет

Пользователи ждут
чат отвечает за десятки секунд

Очередь не знает, сколько займёт запрос



«Занятый консьюмер» знаком всем. Специфика LLM - длину ответа не предскажешь: даже одинаковые запросы держат воркер по-разному.

А как же автоскейлинг GPU?

Автоскейлинг

ожидание: добавим GPU
и пик рассосётся

Ожидание: GPU capacity включается как обычный web autoscale.

А как же автоскейлинг GPU?

Часто нужных GPU у провайдеров физически нет.

Среднее время поставки - от 3 дней до 2 месяцев.

А как же автоскейлинг GPU?

«пошёл доставать GPU»

(c) Саша DevOps

секунды

пик уже случился
чат уже ждёт

А как же автоскейлинг GPU?

«пошёл доставать GPU»

(c) Саша DevOps

секунды
пик уже случился
чат уже ждёт

→

минуты
модель грузится
cache холодный

А как же автоскейлинг GPU?

«пошёл доставать GPU»

(c) Саша DevOps

секунды
пик уже случился
чат уже ждёт

→

минуты
модель грузится
cache холодный

→

часы
квоты, железо,
сетап окружения

Вывод для платформы: держать резерв, отделять тяжёлые классы нагрузки и решать судьбу запроса до GPU.

Сначала разделяем пулы GPU-серверов

Раз пик не доскейлить - работаем со своим флотом GPU: делим его на пулы по профилю нагрузки.

Быстрый интерактив

- TTFT (время до 1-го токена), p95
- короткий output
- резерв важнее утилизации

Сначала разделяем пулы GPU-серверов

Раз пик не доскейлить - работаем со своим флотом GPU: делим его на пулы по профилю нагрузки.

Быстрый интерактив

- TTFT (время до 1-го токена), p95
- короткий output
- резерв важнее утилизации

Длинный контекст

- KV-cache и память
- меньше параллельность
- нужны свои лимиты

Сначала разделяем пулы GPU-серверов

Раз пик не доскейлить - работаем со своим флотом GPU: делим его на пулы по профилю нагрузки.

Быстрый интерактив

- TTFT (время до 1-го токена), p95
- короткий output
- резерв важнее утилизации

Длинный контекст

- KV-cache и память
- меньше параллельность
- нужны свои лимиты

Тяжёлые агенты

tool-вызовы · длинный output · можно ждать дольше

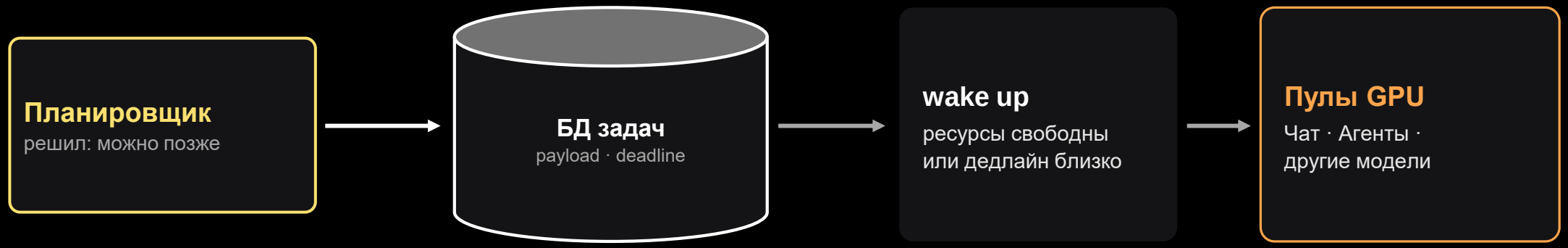
GPU-пулы разделяют профили нагрузки

Пул почти полон - ответы замедляются втрое

Планировщик решает до GPU

Планировщик решает судьбу запроса ДО GPU: сейчас · в очередь · отложить.

Отложенная задача ждёт ресурсов или дедлайна



Отложенные задачи заполняют утилизацию

Коридор утилизации: 60-80 %

ниже коридора
карта оплачена помесячно и простаивает - жжём деньги

рабочий коридор
пул окупается, резерв на пики остаётся

выше коридора
растёт очередь - жжём p95 интерактива

Цепочка: строгий p95 → нужен запас прочности → железо простаивает → дороже каждый ответ.

КЕЙС 3 ИЗ 5

Десять команд, один кластер

Команды ходят к моделям напрямую. Расход растёт, виноватых нет.

Кто съел бюджет CRM?

Ночной скрипт QA
забил очередь батча своими прогонами

Тесты промптов
команда гоняет эксперименты через прод

Расход вырос
непонятно, кто сжёг бюджет

Десять команд в одном кластере без учёта - это общая кредитка без выписки.

Где стоит AI Gateway / LiteLLM

На этом слое запрос получает владельца, бюджет и trace до того, как попадёт в дорогую часть платформы.

Что делает Gateway

доступы
ключи и права по командам

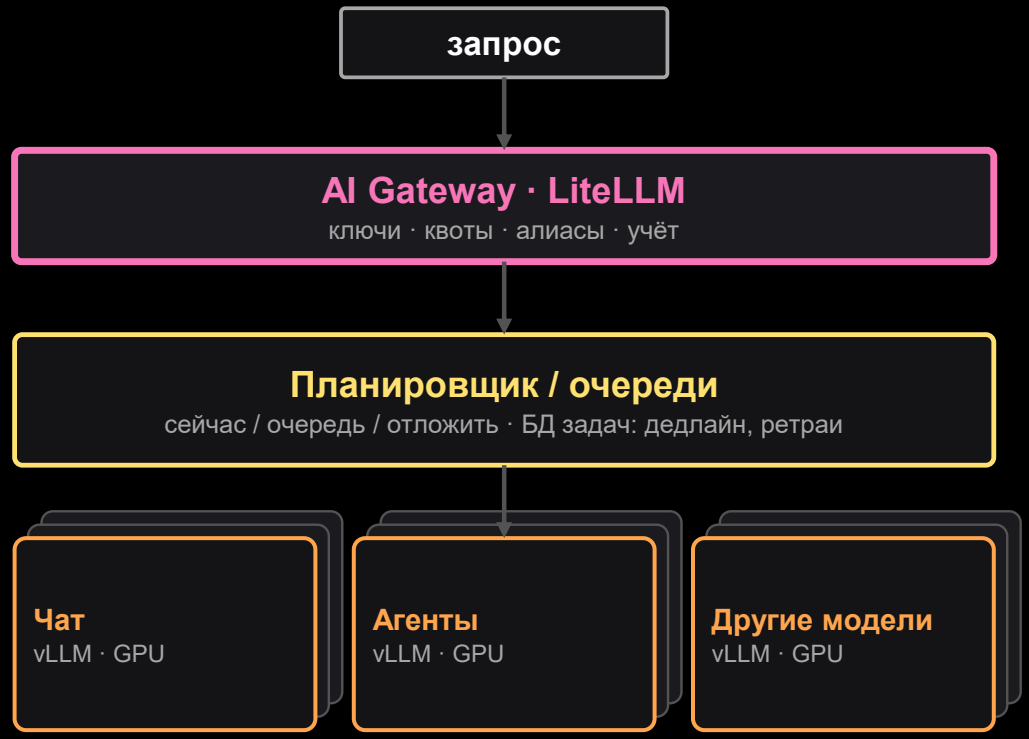
квоты
лимиты на сценарий и команду

usage и trace
кто, куда, сколько потратил

aliases
модель меняется без релиза продукта

FinOps начинается не с Grafana, а с того, что у каждого запроса есть владелец.

Gateway - первый блок AI-контура



КЕЙС 4 ИЗ 5

Очень дорогой «привет»

Появилась дорогая агентская ветка - и туда поехали бытовые запросы.

карта доклада

Обычный чат - один прогон модели

Обычный чат - один прогон модели. Агент выполняет цикл запросов, пока не закончит задачу.

ЧАТ С АССИСТЕНТОМ

Пользователь: «привет, как дела?»

AI: Привет, как ты себя чувствуешь?

прогонов модели: 1

Один «привет» = десятки проходов модели

Обычный чат - один прогон модели. Агент выполняет цикл запросов, пока не закончит задачу.

ДИАЛОГ С АГЕНТОМ

Пользователь: «привет, как дела?»

Агент: вызов: поиск по базе +1 прогон модели

прогонов модели: 1

Один «привет» = десятки проходов модели

Обычный чат - один прогон модели. Агент выполняет цикл запросов, пока не закончит задачу.

ДИАЛОГ С АГЕНТОМ

Пользователь: «привет, как дела?»

Агент: вызов: поиск по базе +1 прогон модели

Агент: вызов: просмотр карточек CRM +1 прогон модели

прогонов модели: 2

Один «привет» = десятки проходов модели

Обычный чат - один прогон модели. Агент выполняет цикл запросов, пока не закончит задачу.

ДИАЛОГ С АГЕНТОМ

Пользователь: «привет, как дела?»	
Агент: вызов: поиск по базе	+1 прогон модели
Агент: вызов: просмотр карточек CRM	+1 прогон модели
Агент: вызов: размышления о бытие	+1 прогон модели

прогонов модели: 3

Один «привет» = десятки проходов модели

Обычный чат - один прогон модели. Агент выполняет цикл запросов, пока не закончит задачу.

ДИАЛОГ С АГЕНТОМ

Пользователь: «привет, как дела?»	
Агент: вызов: поиск по базе	+1 прогон модели
Агент: вызов: просмотр карточек CRM	+1 прогон модели
Агент: вызов: размышления о бытие	+1 прогон модели
Агент: …ещё несколько ходов	+1 прогон модели

прогонов модели: 4

Один «привет» = десятки проходов модели

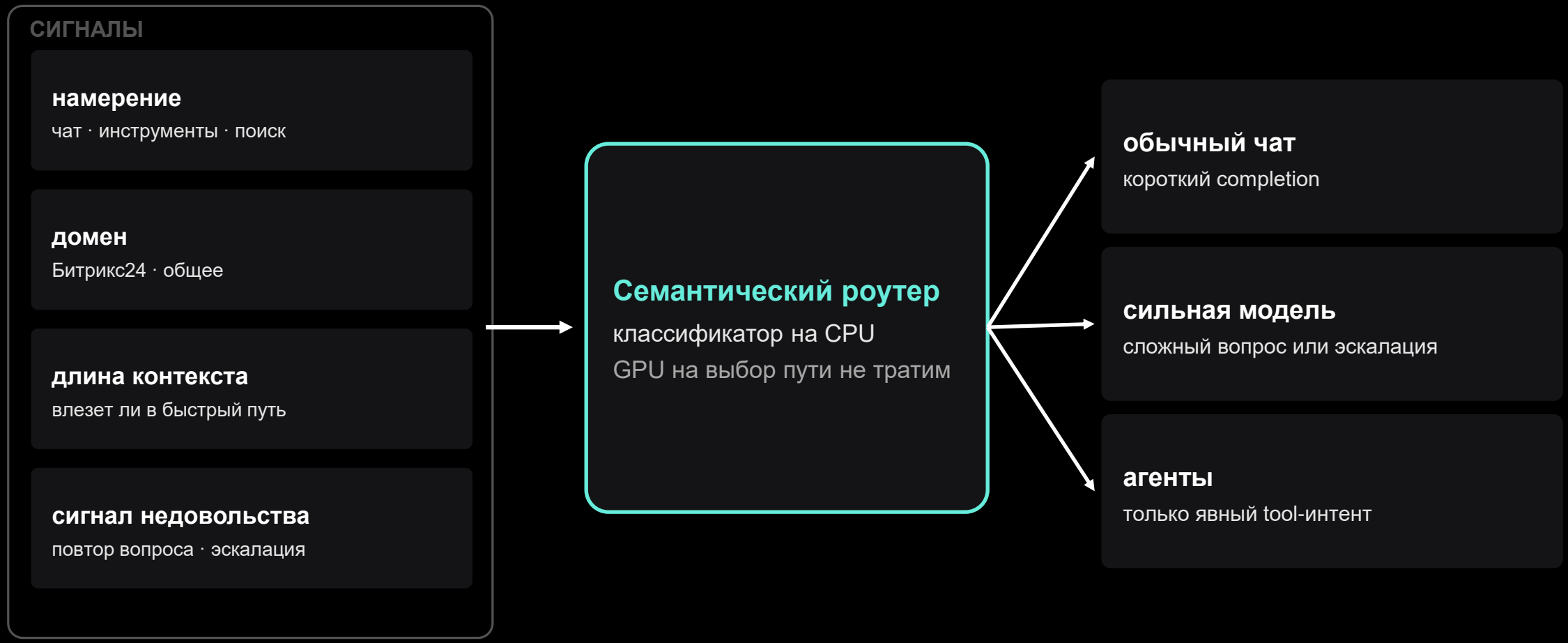
Обычный чат - один прогон модели. Агент выполняет цикл запросов, пока не закончит задачу.

ДИАЛОГ С АГЕНТОМ

Пользователь: «привет, как дела?»	
Агент: вызов: поиск по базе	+1 прогон модели
Агент: вызов: просмотр карточек CRM	+1 прогон модели
Агент: вызов: размышления о бытие	+1 прогон модели
Агент: …ещё несколько ходов	+1 прогон модели
Агент: финальный ответ	

≈ десятки прогонов модели

Семантический роутер выбирает путь до GPU

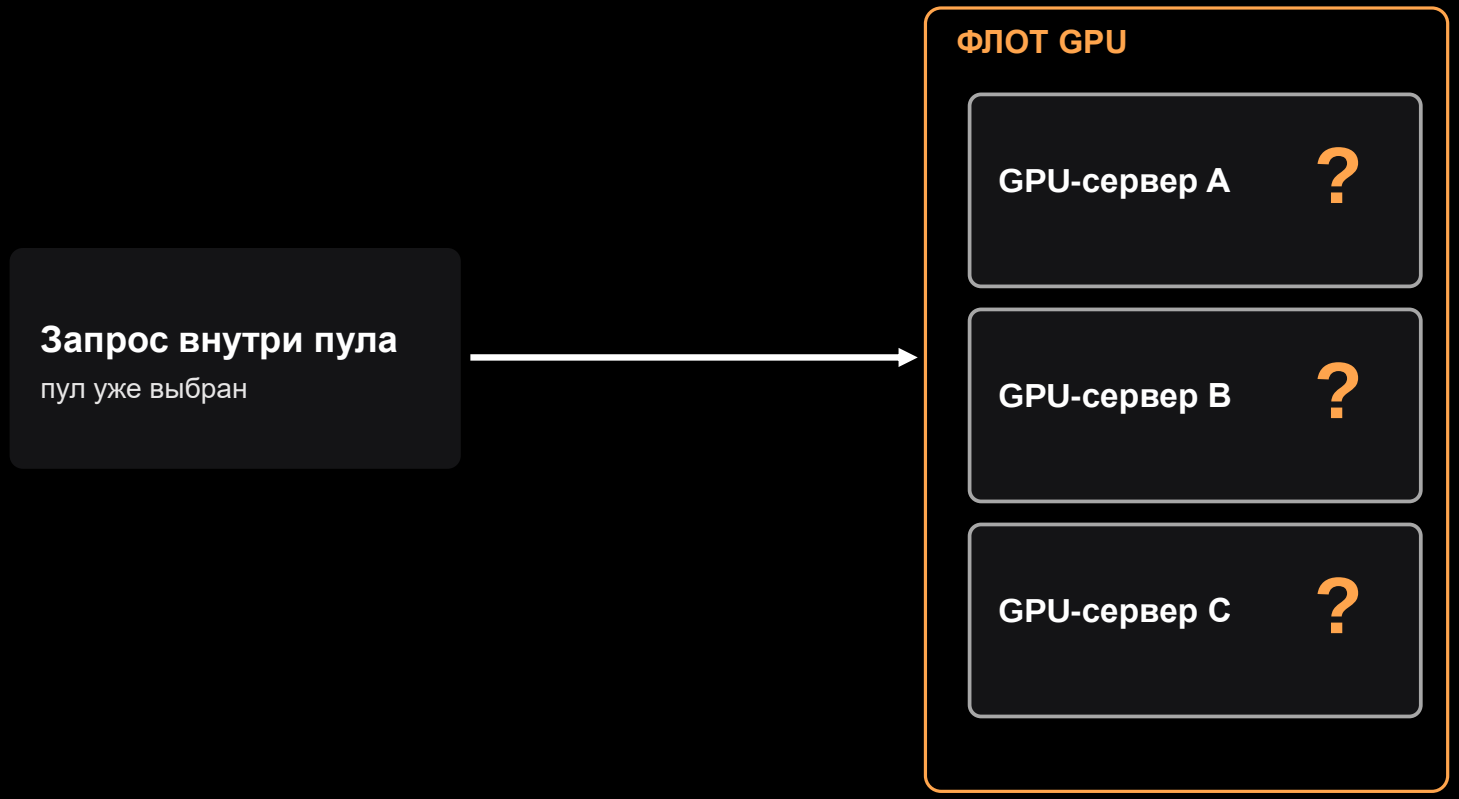


Семантический роутер выбирает путь до GPU

Пул выбрали. А сервер - какой?

В пуле десятки серверов, и они не одинаковые.

Какой GPU-сервер подходит под запрос?



Какая метрика порой важнее нагрузки?

Есть неочевидный сигнал - и он решает.

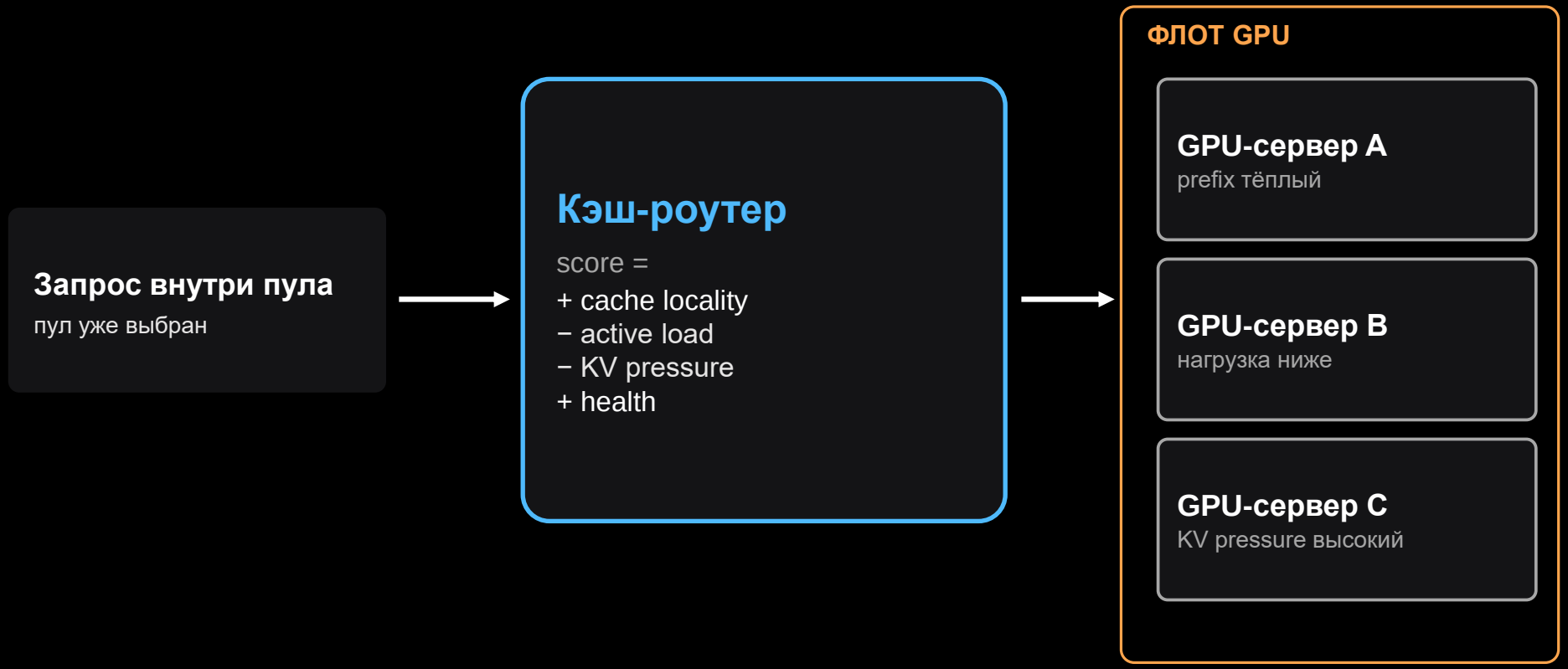
Prefix cache: на чём мы экономим

Общую часть запроса не считаем заново - пока не меняется начало промпта.

1	**Суммаризируй встречу клиента. Верни задачи и риски.**	префикс кэш пустой
2	**Суммаризируй встречу клиента. Верни задачи и дедлайны.**	префикс кэш тёплый
3	**Суммаризируй переписку клиента. Верни задачи и дедлайны.**	префикс кэш: hit rate ~20%

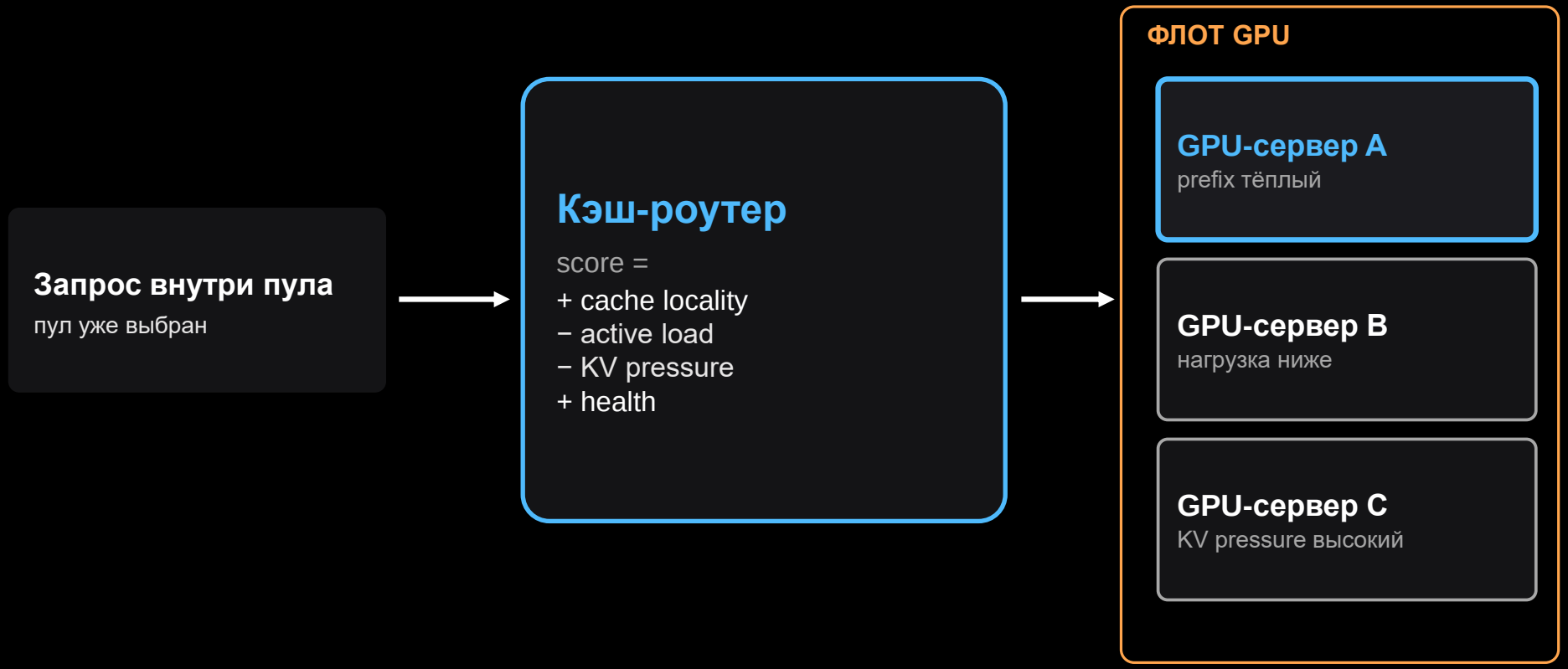
≈52% входных токенов платформы - из кэша: агентская ветка ~91%, массовый чат ~44%, follow-up видеозвонка ~8%.

Кэш-роутер выбирает сервер с тёплым кэшем



Тот же запрос едет на сервер, где префикс уже посчитан - экономим prefill, а не считаем заново.

Кэш-роутер выбирает сервер с тёплым кэшем



Тот же запрос едет на сервер, где префикс уже посчитан - экономим prefill, а не считаем заново.

КЕЙС 5 ИЗ 5

Метрики зелёные, ответы ломаются

Все инфраструктурные метрики зелёные. Ответы модели - уже другие.

FP8 KV-cache выглядит бесплатной оптимизацией

Включили на тестовом сервере: память, throughput и latency стали лучше.

Память лучше
KV-cache компактнее

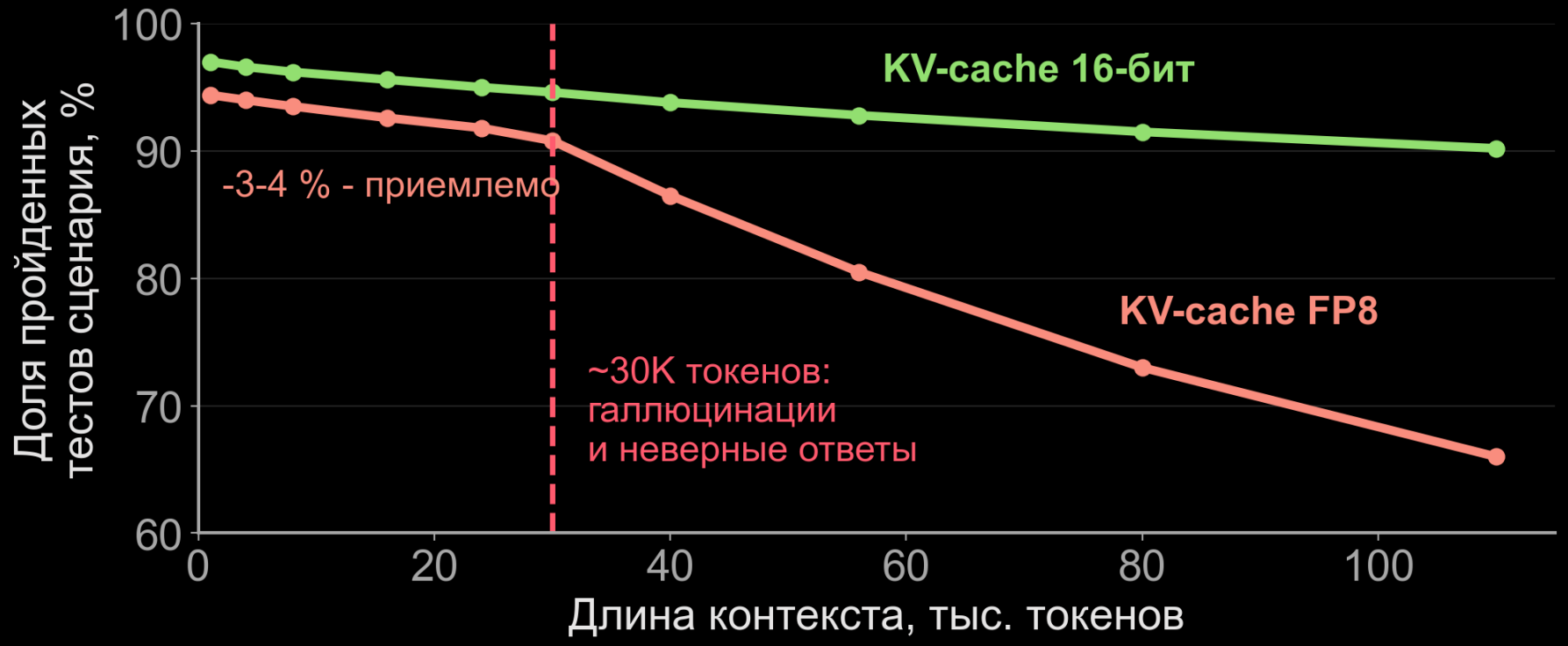
Параллелизм выше
больше запросов на карту

Latency лучше
p95 на тесте улучшился

Ручные проверки проходят
ответы выглядят нормально

На этом этапе мы почти готовы катить в production.

Качество падает после 30K токенов



Поймал follow-up прогон на длинном контексте (>30K), которого не было в ручных проверках.

Любое изменение serving-системы запускает Quality Gate

hash образа vLLM	мог измениться парсер ответа, чат-шаблон или зависимость

Любое изменение serving-системы запускает Quality Gate

hash образа vLLM	мог измениться парсер ответа, чат-шаблон или зависимость
параметры запуска vLLM	kv-cache dtype · quantization · max-model-len - меняют ответ

Любое изменение serving-системы запускает Quality Gate

hash образа vLLM	мог измениться парсер ответа, чат-шаблон или зависимость
параметры запуска vLLM	kv-cache dtype · quantization · max-model-len - меняют ответ
сигнал семантического роутера	новая или изменённая политика

Любое изменение serving-системы запускает Quality Gate

hash образа vLLM	мог измениться парсер ответа, чат-шаблон или зависимость
параметры запуска vLLM	kv-cache dtype · quantization · max-model-len - меняют ответ
сигнал семантического роутера	новая или изменённая политика
параметры кэш-роутера	скоринг и веса серверов изменились

Quality Gate: как устроен прогон качества

Есть Quality Gate - можно менять всё остальное.
Без него любой тюнинг - рулетка.

изменение промпта · модели · квантование · параметры vLLM · routing · cache · runtime version

Что выросло из пяти кейсов

Ради чего: контроль, который стал нашим

Эмоции в транскрибации
почти бесплатно; на рынке - отдельные API

Ночные батчи
ночь оплачена арендой; приоритеты - наши

Настройки под сценарий
пул, TTFT, лимиты контекста - под задачу

Свои приоритеты и цены
квоты, приоритет пикам, тарифы - наши

Готовы ли вы к self-hosted?

0

Есть внешние ограничения по данным?

если их много - MaaS fallback не опция

Готовы ли вы к self-hosted?

0 **Есть внешние ограничения по данным?**

если их много - MaaS fallback не опция

1 **Сценарии тянут модели ~30B-класса?**

если нужна лучшая модель рынка завтра - MaaS быстрее

Готовы ли вы к self-hosted?

0 **Есть внешние ограничения по данным?**

если их много - MaaS fallback не опция

1 **Сценарии тянут модели ~30B-класса?**

если нужна лучшая модель рынка завтра - MaaS быстрее

2 **Нагрузка займёт оплаченные GPU?**

пики, ночные batch-окна, резерв 60-80 %

Готовы ли вы к self-hosted?

0 **Есть внешние ограничения по данным?**

если их много - MaaS fallback не опция

1 **Сценарии тянут модели ~30B-класса?**

если нужна лучшая модель рынка завтра - MaaS быстрее

2 **Нагрузка займёт оплаченные GPU?**

пики, ночные batch-окна, резерв 60-80 %

3 **Есть ли экспертиза и специалисты?**

MLOps / LLMOps и т.д.

Готовы ли вы к self-hosted?

0 **Есть внешние ограничения по данным?**

если их много - MaaS fallback не опция

1 **Сценарии тянут модели ~30B-класса?**

если нужна лучшая модель рынка завтра - MaaS быстрее

2 **Нагрузка займёт оплаченные GPU?**

пики, ночные batch-окна, резерв 60-80 %

3 **Есть ли экспертиза и специалисты?**

MLOps / LLMOps и т.д.

4 **Есть ли достаточный Quality Gate?**

свои evals, не сторонние бенчмарки

В MaaS всё между запросом и ответом зашито в цену токена.
В self-hosted - в людей, процессы и инфраструктурные решения.

Спасибо!

Сергей Нотевский

AI Platform Lead · Битрикс24

Tg: @sergeinotevskii