Корпоративная арифметика AI-трансформации: как не сжечь на токенах, подписках и серверах всю экономию от сокращения ФОТ

CEO и финдиректору корпорации надо немного понимать технику и экономику AI: почему тяжелая разработка сжигает 95 % токенов компании, как пройти яму двойных затрат и когда окупается собственный сервер.

Коротко. CEO и финдиректору корпорации надо немного понимать технику и экономику AI. Покупка подписок и облачных токенов без жесткого финансового плана увеличит ИТ-бюджет на 5-7 %. При старте трансформации бизнес сталкивается с четырьмя «демонами»: иллюзией дешевизны кода, ямой двойных затрат, ценовым капканом провайдеров и рисками доступности плюс КИИ. По моей модели почти 95 % всех токенов компании сжигает тяжелая разработка, где нужны сильные модели и большой контекст. Мое решение для вас - модель «Двойной компьют»: тяжелую разработку перевести на собственный купленный серверный узел, а «мелочевку» + операционные процессы отдавать облачным провайдерам. Эта статья - результат моей работы, дополненный исправлениями по комментариям рецензентов, коллег из индустрии. В статье мысль разверну, но статья скорее для реалий РФ 2026. В других странах ситуация с доступностью мировых сервисов, как вы знаете, намного лучше, там при наличии законодательства можно попробовать найти сверхдешевые облачные решения.

Вопрос на 100 миллионов: реакция на прошлую статью и страх бизнеса

После публикации предыдущей статьи о сокращении до 50 % ИТ-штата к 2028 году коллеги из правлений банков и крупных холдингов стали писать мне в личку. Если собрать их вопросы в один, получится примерно так:

«Сабынин, прогноз твой красивый и хайповый. Но ты сам вообще открывал корпоративные счета за токены и подписки? Многие уже начали с этим «играться» и знают цену «ai-игрушки». Если я завтра выдам каждому сотруднику по условному безлимиту, а оптимизации ФОТ не добьюсь - я не сэкономлю, а тупо потеряю контроль над расходами, просадив финрез. ФОТ, как обычно, остается на месте, а сверху плюс прилипнет счет за облако на 15-20 миллионов рублей в месяц. И где здесь твоя экономика? Кто готов «руку на отсечение» дать, что все получится по твоему плану?»

И да, вопрос абсолютно справедливый. Это реальный управленческий страх, и под ним лежит практика ИТ управления.

Если проводить AI-трансформацию как бесплатную раздачу удобных инструментов сотрудникам (так, например делала Meta), никакого снижения операционных расходов не произойдет. В Meta сотрудники соревновались, кто сожжет больше токенов, пока расходы не пошли к миллиардам и Meta не ввела лимиты. По данным McKinsey за 2026 год, 93 % опрошенных компаний уже превысили свои бюджеты на AI. На мой взгляд, вычисления «сожрут» всю операционную маржу.

Чтобы не «сжечь» деньги акционеров, надо сразу разделить комфортные эксперименты для сотрудников и холодную корпоративную арифметику под российские реалии.

На пути трансформации вас ждут, на мой взгляд, четыре «демона» расходов AI-трансформации. Разберем каждого.

Демон первый: иллюзия дешевизны или почему тяжелая разработка пожирает 94.6 % токенов

Допустим, у вас компания (банк топ 40) на 2 500 человек, из которых примерно 500 - это ИТ (развитие и саппорт). Цифры примерные, но средние. В любой корпорации внедрение AI начинается с офисных восторгов: маркетологи генерируют слоганы и сотни креативов, юристы вычитывают договоры, HR настраивает чат-ботов, фильтруя кандидатов. Руководство видит тысячи запросов в день в AI и делает вывод: «AI уже охватил компанию стихийно, пора бы закупать корпоративный тариф». И это ошибка.

Прежде чем смотреть в расчеты, три базовых термина «на пальцах», без которых финансовому директору и CEO невозможно контролировать расходы:

- **Компьют (Compute)** - это физические вычислительные «мозги»: серверы, специализированные тензорные ускорители (видеокарты) и сверхбыстрая память VRAM. Они круглосуточно считают миллиарды матричных операций, чтобы мы видели сгенерированный текст или код. В облаке вы платите за чужой компьют - за те мощности, которые купил провайдер и сдает вам в аренду с наценкой. При собственном узле - вы покупаете сервер с ускорителями один раз и ставите его в стойку ЦОД.
- **Контекст (Context window)** - это «оперативная память» AI-агента в рамках конкретного запроса: сколько текста и кода модель может одновременно удержать в голове и осмыслить, считается в токенах. Для офисного чат-бота юриста или HR хватает 8 000–16 000 токенов. Топовые американские модели заявляют от 256k до 1M токенов, но для российских реалий и суверенных дата-центров аппетиты придется понизить: локальный сервер стабильно держит честные 128k–256k на команду ядра разработки, и этого с запасом хватает под сложные задачи.
- **Токен (Token)** - это единица измерения данных и тарификации, за которую вам и выставляют счет. По сути, это слог или кусочек слова: в русском тексте 1 токен равен примерно 3-4 символам (в слове «трансформация» - три токена). Когда модель читает входящий запрос, монорепозиторий или документ - вы платите за входящие токены (Input). Когда формулирует ответ или генерирует код - за исходящие (Output), которые у коммерческих провайдеров стоят в 3-4 раза дороже. В масштабах компании счет идет на миллиарды токенов ежемесячно.

В прошлой статье я говорил банальную истину, что все корпоративное ИТ делится на два фундаментальных контура: **Развитие** и **Саппорт**. Если наложить эту структуру на математику потребления токенов, получается показательная картина:

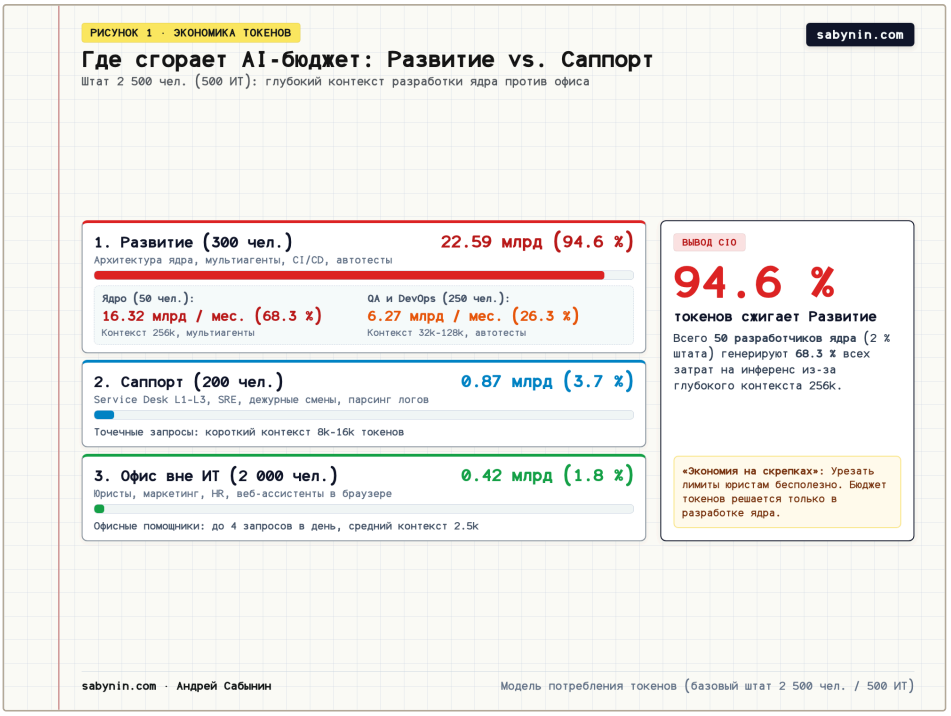


Рисунок 1. Где сгорает AI-бюджет компании: Развитие vs. Саппорт

Сводные данные по компании (детальная модель с формулами по каждому подразделению, нормативами запросов и тарифами Input/Output выложена в файле ai-infrastructure-spec-2026.xlsx, Лист 1):

Контур и роль	Штат (доля)	Профиль нагрузки (21 раб. день)	Средний контекст AI агента (вход + выход)	Объем в месяц	Доля в токенах
1. Развитие: Core и Архитектура (мультиагентные цепочки)	**50 чел.** (2 %)	60 запусков/день	256k + 3k токенов	**16.32 млрд**	**68.3 %**
2. Развитие: QA, Аналитика, DevOps (автотесты, CI/CD, требования)	**250 чел.** (10 %)	25 вызовов/день	46k + 1.8k токенов	**6.27 млрд**	**26.3 %**
ИТОГО РАЗВИТИЕ (Change)	**300 чел.** (12 %)	—	**Глубокий контекст**	**22.59 млрд**	**94.6 %**
3. Саппорт (Run) (L1–L3, SRE, разбор логов и инцидентов)	**200 чел.** (8 %)	20 вызовов/день	9.5k + 0.9k токенов	**0.87 млрд**	**3.7 %**
4. Офис вне ИТ (маркетинг, продажи, юристы, HR — чат в браузере)	**2 000 чел.** (80 %)	4 запроса/день	2k + 0.5k токенов	**0.42 млрд**	**1.8 %**
ВСЕГО ПО КОМПАНИИ	**2 500 чел.** (100 %)	**446 тыс. циклов/ мес.**	—	**23.88 млрд**	**100 %**

Суммарно по компании набегает **23.88 млрд токенов в месяц**, из которых контур Развития сжигает **22.59 млрд токенов (94.6 %)**, а весь Саппорт и двухтысячный офис вместе взятые - лишь 1.29 млрд (5.4 %). Бороться за экономию, урезая лимиты офису - классическая экономия на скрепках. Управлять надо разработкой и ее эффективностью в первую очередь.

Это нижняя граница, скромная эксплуатация. У активных пользователей агентов объем в разы выше, поэтому дальше я считаю и вариант ×2, 45 млрд токенов в месяц. Кому-то и эта оценка покажется заниженной, но давайте будем реалистами: в реалиях РФ дешевые по сравнению с API подписки американских лабораторий недоступны, и работать приходится с тем, что доступно.

При этом в разработке слабая AI-модель – катастрофа, поверьте. В сложном цикле Развития слабая модель без глубокого контекста не держит архитектуру, лепит костыли и спагетти-код (потом к вам еще придут аутсорс-вендоры разгребать этот «завал» за тройной прайс). Именно поэтому разработчики тайно через VPN заходят по вашим рабочим задачам в американский Claude и Codex, создавая риски безопасности – просто эти модели реально справляются со сложными задачами.

Инженерам осенью 26 года бескомпромиссно нужен честный контекст AI агентов 128k-256k токенов на сильных флагманских моделях (DeepSeek-V4, GLM 5.3, Kimi K3.1).

Откуда берется счет на **9-14 миллионов рублей ежемесячно** в облаке? Тарифы облачных провайдеров на коммерческие флагманские модели при глубоком контексте (>128k) с учетом повышающего коэффициента составляют в среднем 400–600 ₽ за 1 млн токенов в общем потоке вычислений (ссылки на тарифы в конце статьи). При объеме 22.59 млрд токенов в месяц в контуре Развития прямой счет за облачный API составляет **9-14 млн рублей в месяц** (110-160 млн рублей в год) — без гарантий доступности и без учета пиковых нагрузок. Финдиректор точно не обрадуется росту затрат, хотя я иногда думаю, что это копейки на фоне эпохи костов на аутстафф середины 2010-х.

Демон второй: почему важен тайминг трансформации и Трудовой кодекс (ловушка двойных затрат)

В презентациях консультантов все гладко: внедрили AI, продуктивность выросла, штат сократился на 30-50 %, затраты упали. В жизни генеральный и финансовый директора попадают в классическую «яму двойных затрат».

Давайте пример. Периметр: ИТ банка или холдинга на 500 человек. При средней ставке специалиста в 420 000 рублей базовый ФОТ составляет условно **210 миллионов рублей ежемесячно**.

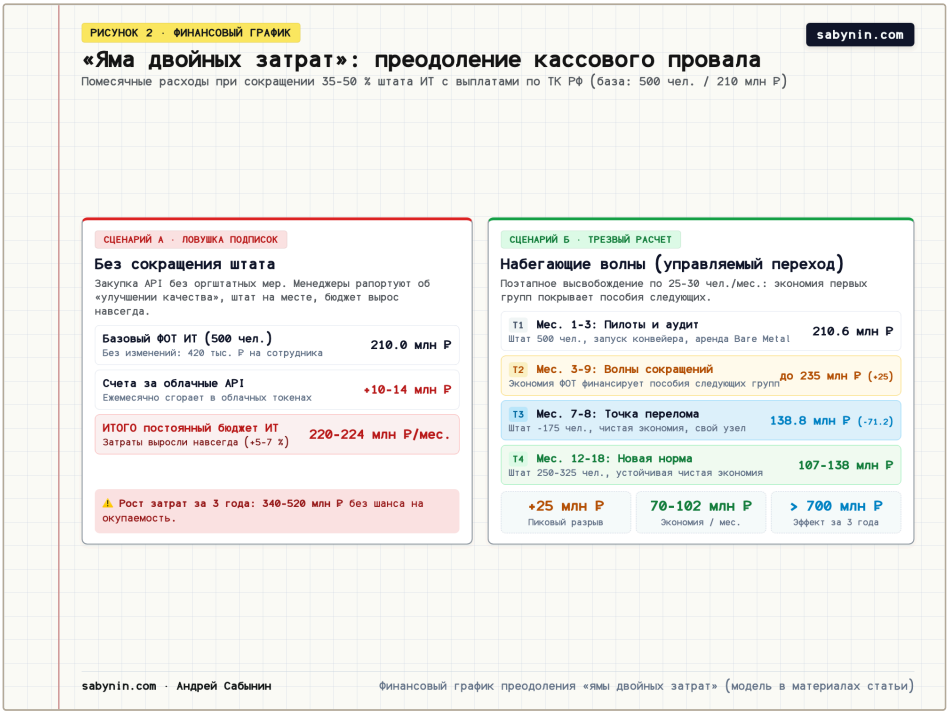


Рисунок 2. «Яма двойных затрат»: как пройти кассовый провал при оптимизации ФОТ

Сравнение трех сценариев прохождения трансформации (помесячный график оптимизации ФОТ, выплат по ТК РФ и модели самофинансирования волн сведен в файле ai-infrastructure-spec-2026.xlsx на Листе

3 «Модель ФОТ и волны»):

Параметр (база: 500 ИТ, ФОТ 210 млн ₽/мес.)	Сценарий А: «Ловушка подписок»	Сценарий Б1: «Два залповых удара» (типичная ошибка)	Сценарий Б2: «Набегающие волны + лизинг» (целевой)
Механика по оптимизации	Штат не сокращается («люди работают качественнее»)	Увольнение 175 чел. двумя волнами в мес. 4 и 6	По 25–30 чел./мес. с 3-го по 9-й месяц (самофинансирование)
Инфраструктура	Покупка облачных подписок и API всем подряд	Облако до 7-го мес., разовая покупка сервера (84 млн ₽)	День 1 — аренда Bare Metal (0.6 млн), далее 3-летний лизинг узла
Пиковый кассовый разрыв	**+10–14 млн ₽/мес.** навсегда	**+68 млн ₽** в 4-м мес. (бюджет прыгает до 278 млн ₽)	**+25 млн ₽** (пособия платятся из экономии прошлых волн)
Юридический и HR-риск	Нулевой (деньги просто сгорают)	Критический: массовое сокращение, проверки ГИТ, уход сеньоров	Низкий: ниже порогов ГИТ, 50 сеньоров защищены программой удержания
Ежемесячный бюджет через год	**220–224 млн ₽/мес.** (+5–7 % к базе)	**138.8 млн ₽/мес.** (если бизнес переживет шок)	**108–138 млн ₽/мес.** (экономия 70–102 млн ₽/мес.)
Изменение затрат за 3 года	**Рост затрат на 340-520 млн ₽**	**Снижение затрат на 650 млн ₽** (с высоким риском остановки поставки)	**Снижение затрат более чем на 700 млн ₽** (без кредитов под ставку ЦБ 14 %)

Главный секрет перехода — **набегающие волны и самофинансирование**: штат сокращается не залпом, а по 25–30 человек ежемесячно с 3-го по 9-й месяц по мере перевода каждой команды на конвейер ИИ. Экономия ФОТ от первых когорт (уже с 5-го месяца это ~31 млн ₽/мес.) постепенно «оплачивает» выходные пособия следующих, сжимая кассовый разрыв с +68 млн до управляемых +25 млн ₽ без дорогих кредитов под ставку ЦБ 14%.

А чтобы 50 ключевых архитекторов и ведущих инженеров не уволились при первых слухах, их до старта выделяют в защищенную категорию главных ИИ-архитекторов. Премии руководителей команд отвязываются от численности подчиненных и привязываются к себестоимости поставки доработки в бой. Направление 10–15 % сэкономленного ФОТ в фонд удержания (+25–30 % к окладу оставшихся) превращает сопротивление во внутреннюю селекцию силами самих команд.

Демон третий: зависимость от чужой инфраструктуры или как не попасть в ценовой капкан?

Сейчас на рынке разворачивается классическая платформная экспансия: провайдеры **«подсаживают» корпоративных клиентов на свои API, как драгдилеры**.

Наглядный пример, вспомните такси в 2010-х: поездки стоили копейки, водителям доплачивали за счет венчура. Как только таксопарки закрылись, а пользователи привыкли жать кнопку - тарифы взлетели, а комиссия всем известного сервиса выросла до 30 %, водители и клиенты жалуются, а уже поздно.

В AI, возможно, будет то же самое. Вендоры субсидируют токены, пока вы перестраиваете процессы и увольняете людей, «подсаживаясь на иглу». А когда компания завязала критические процессы на внешний API без собственных вычислительных мощностей, начнется банальный ценовой шантаж. Сказать «мы уходим» будет

тяжело: нанять людей обратно сложно, да и институциональная память будет частично утеряна. Повышение цен тогда просто переложит экономию ФОТ в карман облачного монополиста. Кстати, именно поэтому, я говорил ранее в первой статье, надо сохранять лучшую экспертизу даже при оптимизации.

Демон четвертый: бесперебойность операционной деятельности и соблюдение законов

Очевидно, если вы работаете с облачным провайдером, требуйте гарантированного SLA 99.99 % на задержку ответа. Особенно это актуально для текущих форс-мажоров с датацентрами. Если модели внедрены в живой операционный процесс (выдача кредита, антифрод), а внешний сервис завис - бизнес встал.

Поэтому регуляторика на пальцах:

1. **Ни один безопасник ИТ банка не согласует отправку исходного кода критичных систем и скоринга во внешнее SaaS-облако.** В коде могут быть зашиты структуры баз данных, токены и банковская тайна (ст. 26 395-1), а внешние ai-агенты за этим не следят.
2. **К тому же закон об ИТ-аутсорсинге для финансовых организаций (законопроект № 404786-8, Госдума приняла его 14 октября 2026 года, Совет Федерации одобрил)** требует заранее уведомлять ЦБ РФ с заключением операционного аудита и держать системы исполнителя в России, а объекты КИИ подчиняются отдельным требованиям 187-ФЗ.
3. **Утечка кода через личные подписки и VPN в публичную сеть** - прямой повод для предписаний ЦБ, отзыва лицензии и персональной ответственности руководства.

Отдавать критические процессы внешнему API без своего серверного узла - недопустимо. Что же делать?

«Истребитель демонов»: собственный серверный узел и модель двойного контура

Если внешние облака несут такие риски, нужно ли строить дата-центр? Нет. Базовый вычислительный блок для компании на 2 500 человек - **один серверный узел на 8 специализированных ускорителей с 1 ТБ памяти VRAM**. Этого достаточно, чтобы локально развернуть квантованные флагманские модели (DeepSeek-V4, GLM 5.3) и дать разработчикам честный контекст 256k.

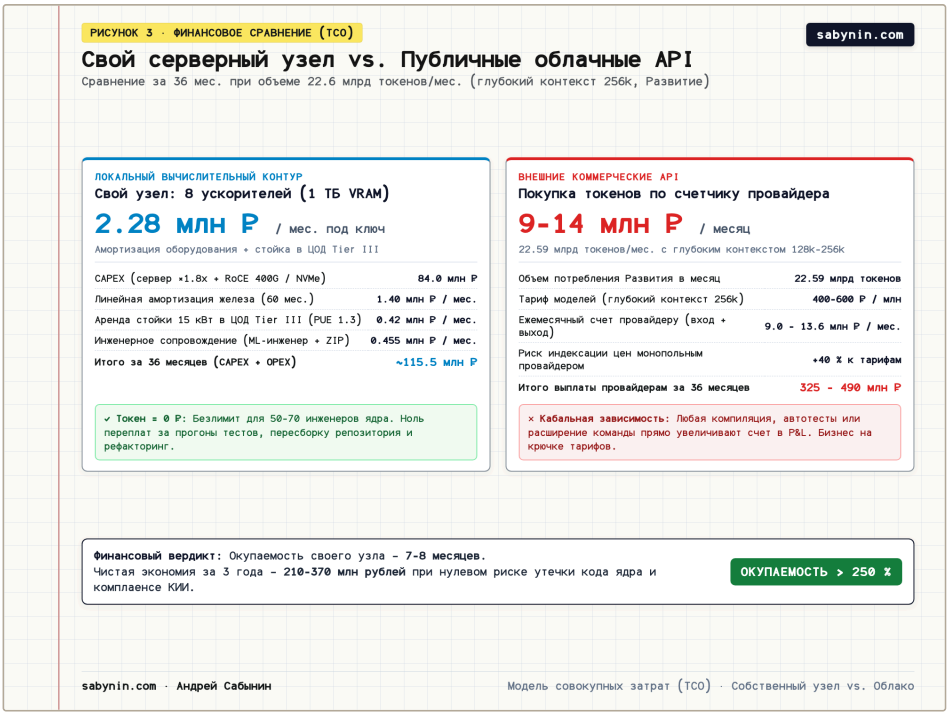


Рисунок 3. Собственный кластер vs Публичные облачные API

Финансовое сравнение собственного серверного узла с коммерческим облачным API:

Показатель (на объем 22.6 млрд токенов/мес.)	Публичный облачный API (по счетчику)	Свой узел: Базовый TCO (покупка, аморт. 5 лет)	Свой узел: Стресс-тест CFO (лизинг 3 года, ставка ЦБ 14 %)
Разовый CAPEX на старте	0 ₽	**84.0 млн ₽** (сервер в РФ 66 млн + сеть RoCE 400G и NVMe 18 млн)	**0 ₽** (вход через аренду 0.6 млн/мес. и перевод в лизинг)
Амортизация / лизинговый платеж	—	**1.40 млн ₽/мес.** (60 мес.)	**3.05 млн ₽/мес.** (36 мес. с учетом стоимости капитала)
Эксплуатация (OPEX в ЦОД Tier III)	Включено в тариф	**0.88 млн ₽/мес.** (стойка 15 кВт 280k, свет 140k, 0.5 FTE 275k, ЗИП и связь 180k)	**0.88 млн ₽/мес.**
ИТОГО ежемесячный платеж (TCO)	**9.0 – 13.6 млн ₽/мес.** (среднее: 11.3 млн ₽)	**2.28 млн ₽/мес.** (фиксировано, безлимит)	**3.93 млн ₽/мес.** (фиксировано, безлимит)
Себестоимость 1 млн токенов	**400 – 600 ₽** (в среднем ~500 ₽)	**101 ₽** (маржинальная цена доп. токена = 0 ₽)	**174 ₽** (в 2–3 раза дешевле облака)
Совокупные затраты за 36 месяцев	**325 – 490 млн ₽** (+ риск роста цен на 40 %)	**115.5 млн ₽** (84 млн CAPEX + 31.5 млн OPEX)	**141.5 млн ₽** (полная выплата лизинга и OPEX)
Чистая экономия за 3 года (только на токенах)	**База сравнения (0 ₽)**	**+210…+370 млн ₽** (окупаемость ~8 мес.)	**+185…+345 млн ₽** (окупаемость с 1-го месяца)

Главное сравнение с облаком

За 2.3 млн рублей в месяц ваш сервер дает **полный безлимит вычислений и свободу выбора моделей**: у локального узла себестоимость дополнительного токена равна нулю. В облачном API объем в 22.6 млрд токенов обходится в **9-14 млн рублей ежемесячно**.

Чистая экономия за 3 года только на разнице между своим сервером и облачными токенами составляет **210-370 млн ₽** (и свыше 1.8 млрд ₽ суммарно с оптимизацией ФОТ!). Даже при жестком финансовом стресс-тесте со ставкой капитала ЦБ 14% собственный узел в 2-3 раза выгоднее облака и окупается за 7-8 месяцев.

А если взять те же модели в облаке и включить кэш?

Справедливый вопрос коллег после публикации: зачем сравнивать свой узел с YandexGPT и GigaChat по полному тарифу, если те же открытые флагманы уже есть в российских облаках, а повторно прочитанный контекст провайдеры считают со скидкой? У агентной разработки до 90 % входящих токенов - кэш: агент на каждом шаге перечитывает один и тот же репозиторий.

Пересчитал по прайсу Cloud.ru на GLM 5.3 (вход 173 ₽, кэш 43 ₽, выход 606 ₽ за 1 млн токенов с НДС) при 90 % кэша:

Показатель	×1: 22.6 млрд токенов/мес. (скромная эксплуатация)	×2: 45.2 млрд токенов/мес. (активная эксплуатация)
Облако, те же флагманы, кэш сработал	**1.5 млн ₽/мес.**	**3.0 млн ₽/мес.**
Облако, те же флагманы, кэш не сработал	4.1 млн ₽/мес.	8.2 млн ₽/мес.
Российские облачные модели по полному тарифу (400-600 ₽)	9.0-13.6 млн ₽/мес.	18.1-27.1 млн ₽/мес.
Свой узел, покупка	**2.3 млн ₽/мес.**	**2.3 млн ₽/мес.** (оценка: один узел справляется за счет кэша)
Свой узел, лизинг	3.9 млн ₽/мес.	3.9 млн ₽/мес.

Вывод честный: при скромной эксплуатации облако с теми же моделями и работающим кэшем дешевле своего узла. Порог, после которого покупка узла выгоднее, - около **34 млрд токенов в месяц**, для лизинга - около 59 млрд. Лизинг против облака с кэшем не экономит, его задача - убрать CAPEX и кассовый разрыв на старте.

Но у кэша два «но». У российских провайдеров он не гарантирован: Yandex включает его «когда возможно», GigaChat требует тот же идентификатор сессии и не публикует срок хранения. А без кэша облако уже при скромной эксплуатации стоит 4 млн в месяц. И главное, код ядра, схемы БД и банковская тайна во внешнее облако все равно не уйдут, об этом демон четвертый.

Поэтому логика двойного контура остается, уточняется только порядок: начинать с облака или аренды сервера, замерить реальный объем токенов за 2-3 месяца и переводить разработку на свой узел, когда объем подходит к порогу или этого требуют ИБ и регулятор. Расчет порога и чувствительность к доле кэша - на Листе 4 таблицы.

А потянет ли один узел 50 разработчиков по факту?

Вопрос вашего технического директора: «У вас в узле 1 ТБ памяти. Если 50 сильных инженеров параллельно загрузят задачи с контекстом 256k - сервер не ляжет?».

Отвечаю: **потянет с запасом**. 50 человек никогда не нажимают «Enter» в одну и ту же миллисекунду. Нагрузка распределена по дню, а архитектуры моделей (DeepSeek-V4, GLM 5.3, Kimi K3.1) держат поток за счет латентного сжатия памяти (MLA/DSA), дедупликации общих префиксов репозитория (RadixAttention) и выгрузки неактивных сессий на быстрые накопители NVMe.

А хватит ли одного узла на 45 млрд токенов в месяц, если его купить? Считаю на пальцах. Рабочий месяц - 21 день по 10 часов, это 756 тыс. секунд. 45.2 млрд токенов на это время - около 60 тыс. токенов в секунду. Звучит много, но около 90 % из них - кэш: агент перечитывает тот же репозиторий, и узел не пересчитывает его заново, а берет готовым из памяти. Считать по-настоящему надо около 6 тыс. токенов в секунду нового входа и около 1.1 тыс. токенов в секунду ответа.

Сколько дает узел? DeepSeek публиковал данные своего сервиса: узел из 8 ускорителей H800 в среднем обрабатывал около 74 тыс. токенов в секунду на входе, больше половины из них из кэша, и около 15 тыс. токенов в секунду на выходе. DeepSeek обрабатывает вход и ответ на разных узлах, а я складываю оба потока на один узел, поэтому беру в три раза меньше, тем более что у нас один узел, а не кластер, и квантованная модель: примерно 10 тыс. токенов нового входа и 5 тыс. токенов ответа в секунду. Нагрузка выходит около 55-60 % по входу и 20-25 % по ответу. Остается запас на пики и на ночные пакетные задачи: прогон тестов, ревью, документацию. При ×10 так уже не выйдет, понадобятся еще узлы.

Провокационный тезис: разработка без отказоустойчивости, а клиентский Саппорт - в дешевое облако

Когда архитекторы приносят смету на 250 миллионов рублей за резервирование внутреннего облака (два узла в разных ЦОД, гео-репликация), задайте циничный вопрос: **а что случится, если сервер внутреннего ИИ для разработчиков зависнет на 45 минут?**

Платежи остановятся? Банкоматы перестанут выдавать наличные? Регулятор отзовет лицензию? **Ничего не случится.** Программисты вздохнут, пойдут налить кофе, обсудят архитектуру у доски или напишут код руками, как делали последние тридцать лет. Тяжелая разработка (контур Развития, 94.6 % токенов) - это внутренний контур, где заказчики привыкли ждать задач месяцами, все равно будет быстрее и намного, чем «вчера».

Оговорка, которую справедливо сделали коллеги: «напишут руками» работает, пока люди есть. Когда штат уже сокращен и производство держится на агентах, простой узла надо считать через операционный риск. Решение недорогое: заранее заключить договоры с оплатой по факту у двух облачных провайдеров с теми же открытыми моделями и поднять лимиты до аварии. Пока узел работает, такой резерв почти ничего не стоит, а при 5 днях простоя в год обходится в 40-70 тыс. ₽ в месяц (Лист 4).

Настоящая непрерывность 24/7 критически важна только в клиентских сервисах и Саппорте:

Архитектурный критерий	Контур 1: Тяжелое Развитие (Change)	Контур 2: Клиентский Саппорт и Офис (Run)
Доля трафика и контекст	**94.6 % всех токенов**, глубокий контекст 128k–256k	**5.4 % всех токенов**, короткие запросы 2k–4k
Критичность простоя на 1 час	**Нулевая для выручки:** программисты пьют кофе или пишут руками, платежи клиентов не встают	**Критическая (24/7):** остановка клиентских ботов, скоринга и колл-центра
Требования ИБ (395-1, 187-ФЗ, 404786-8)	**Абсолютный запрет на облако:** внутри исходный код ядра АБС, схемы БД и банковская тайна	**Допустимо внешнее облако** после фильтрации ПДн на внутреннем DLP-шлюзе
Где размещаем и какие модели	**Свой узел On-Prem** (8 ускорителей, 1 ТБ VRAM): открытые флагманы `DeepSeek-V4`, `GLM 5.3`, `Kimi K3.1`	**Дешевые внешние API** с гео-резервом провайдера: `DeepSeek-V4.1-Flash`, `YandexGPT Lite`, `GigaChat`

Архитектурный критерий	Контур 1: Тяжелое Развитие (Change)	Контур 2: Клиентский Саппорт и Офис (Run)
Схема отказоустойчивости	**Без 2N-кластера** (экономия 250 млн ₽). Страховка: холодный ЗИП в ЦОД (замена за 4 ч) + резервный `AWQ 4-bit` на Bare Metal за 60 мин	**SLA 99.99 % «из коробки»** за счет субсидированной инфраструктуры облачного гиганта
Цена вопроса	**2.28 млн ₽/мес.** (полный безлимит вычислений)	**80 – 120 тыс. ₽/мес.** (жесткие квоты на отдел)

Архитектура «рачительного CIO»

Гибридная модель двойного контура:

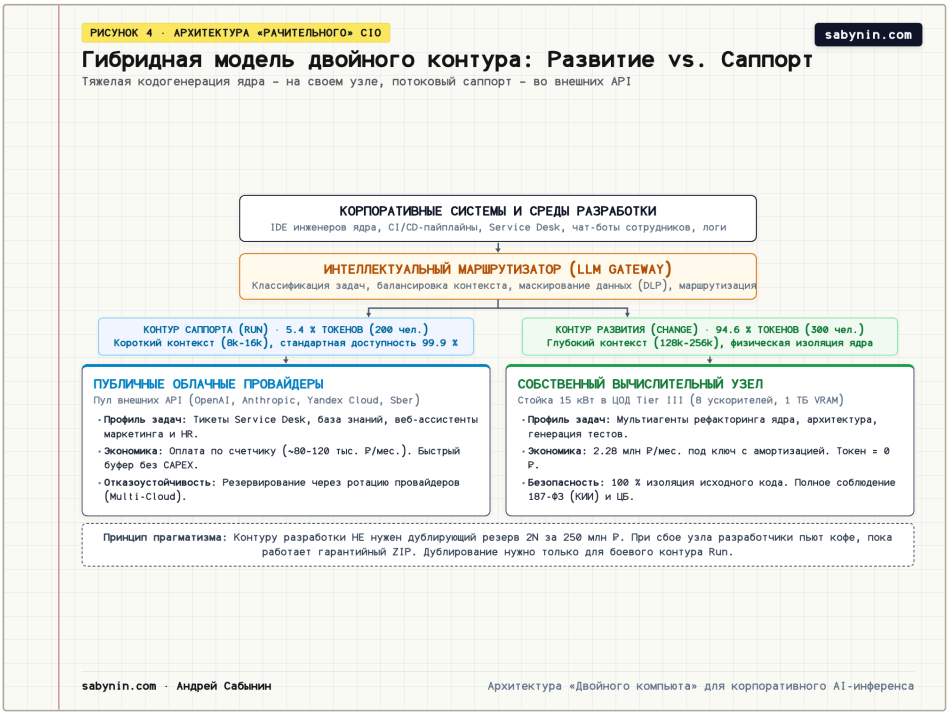


Рисунок 4. Архитектура «рачительного CIO»: модель двойного контура

1. **Единый внутренний AI-шлюз:** все запросы идут через единую корпоративную точку (проверка прав, DLP-маскирование данных, аудит и умная маршрутизация).
2. **Дифференцированные квоты:** офису и клиентским сервисам — жесткие лимиты на легких внешних моделях с SLA 99.99%, отсекая спам.
3. **Изолированный контур ядра:** разработка критичных систем живет на собственном сервере под контролем ИБ, без затрат на избыточный второй кластер.
4. **Инверсия старта и лизинг:** День 1 — аренда выделенного сервера (~600 тыс. ₽/мес.), при масштабировании — перевод узла в 3-летний лизинг.
5. **Мандат на оптимизацию:** доступ к агентам открывается командам строго под обязательство снижения себестоимости фичи и сокращения штата волнами.

Вывод

AI-трансформация - это точно не технологический проект ИТ-директора. Это жесткая финансовая реорганизация компании, и спокойно считается в деньгах.

И если просто натянуть AI-агентов поверх старых раздутых процессов и прежнего штата, вы гарантированно получите ухудшение P&L и, возможно, кабальную зависимость от провайдеров.

Выиграют те руководители, которые с первого дня подойдут к процессу скрупулезно - с карандашом, калькулятором и финансовым директором: рассчитают график набегающих волн, договорятся с ключевым ядром инженеров и построят суверенный вычислительный контур.

Детализированный расчет потребления токенов по подразделениям, методика пересчета контекста и полная аппаратная спецификация серверного узла выложены в файле: **ai-infrastructure-spec-2026.xlsx** (внутри: помесячная разбивка по ролям, формулы тарификации Input/Output, спецификация серверного узла с ценами в Китае и доставкой в РФ, модель оптимизации ФОТ по набегающим волнам, а на Листе 4 - облако с кэшем, порог окупаемости своего узла и резерв на простой).

Источники

- Тарифы Yandex AI Studio: aistudio.yandex.ru
- Тарифы GigaChat API: обзор тарифов Сбера с 1 февраля 2026 года, тарифы для юрлиц, кэшированные токены не тарифицируются
- Тарифы Cloud.ru на открытые модели, включая GLM 5.3: прайс Evolution Foundation Models
- Тарифы MWS GPT: mws.ru
- Meta: внутренний рейтинг расхода токенов среди сотрудников и последующие лимиты: Fortune, The Decoder
- DeepSeek, пропускная способность узла: DeepSeek-V3/R1 Inference System Overview
- McKinsey, перерасход бюджетов на AI: Burning through the AI budget
- Закон об ИТ-аутсорсинге в финсекторе (законопроект № 404786-8): Интерфакс, Право.ru

---



Андрей Сабынин
Экс-член правления банка, CIO, основатель AI-native ИТ-компании. AI-трансформация и снижение стоимости ИТ в банках и финтехе на практике.