

СТРАТЕГИЧЕСКАЯ СТАТЬЯ

Редакционная дата: 24 октября 2026 года

Компьютер, который сам снижает себе оплату

*Почему цена интеллекта может упасть раньше, чем окупится оборудование?**The Token Deflation Trap: When Faster Inference Undermines Compute Payback*

Вступление: процессор может превосходно справляться с задачей, ради которой его купили, и в тот же день уже отставать экономически. Это не загадка о неисправных чипах, а вопрос об отрасли, способной наращивать выпуск интеллекта быстрее, чем она успевает брать за него плату. Мы откроем американскую финансовую отчётность, затем обратимся к алгоритму, который имитирует страничную организацию памяти, а после — к вычислительной инфраструктуре, уже введённой в эксплуатацию в Саудовской Аравии. Возможно, решающая конкуренция идёт не за количеством вычислений, возможных на бумаге, а за той минутой, когда вычисление становится полезным и пригодным для продажи.

1. В бухгалтерских книгах чип живёт шесть лет, а на рынке новый соперник появляется каждый квартал

В годовом отчёте CoreWeave за 2025 год есть короткая фраза, мимо которой легко пройти, как сотруднику мимо расчётного листка без сюрпризов: с января 2023 года компания увеличила расчётный срок полезного использования вычислительного оборудования в своих центрах обработки данных с пяти до шести лет. Согласно раскрытию, решение связано с развитием аппаратной части, совершенствованием программного обеспечения и проектированием центров обработки данных. Это

расчётный бухгалтерский срок, по которому распределяется амортизация (Depreciation), а не обещание, что устройство останется самым прибыльным на протяжении шести лет S01.

5 января 2026 года NVIDIA представила платформу Rubin и заявила, что её интегрированная архитектура способна снизить **себестоимость выпуска токена при инференсе** (Inference Token Cost) до десяти раз по сравнению с Blackwell на отдельных нагрузках. Это заявление поставщика о целевой технической возможности, а не раскрытие фактических отпускных цен и не вывод обо всех моделях и приложениях S03.

Затем вышла отчётность за первое полугодие 2026 года: CoreWeave сообщила, что на конец июня объём оставшихся обязательств к исполнению (Remaining Performance Obligations; RPO) достиг **103,7 млрд долларов**. Цифра впечатляет, но ещё важнее её определение: это договорная стоимость услуг, которые ещё не оказаны, рассчитанная по правилам учёта и подверженная возможному влиянию задержек и скидок. Это не деньги в кассе и не гарантированная чистая прибыль. Компания ожидала признать около 41% этой суммы в следующие 24 месяца, а ещё 39% — в последующие 24 месяца S02.

Разговор об искусственном интеллекте обычно начинают с названия чипа и объёма его памяти. Но папка, где рядом лежат «шесть лет», «десятикратное снижение стоимости токена» и «103,7 млрд будущих услуг», ставит перед нами более занятную и опасную задачу: **какое из этих времён должно выиграть гонку?** Срок амортизации? Срок выхода нового поколения техники? Срок контрактов, защищающих денежные потоки? Или скорость, с которой меняется цена, приемлемая для клиента?

Ни одна из этих цифр не доказывает крах рынка и не показывает, что конкретная компания попала в ловушку. Напротив, долгосрочные контракты — веская причина не сводить всё к истории о «процессоре, который обесценивается сразу после выхода преемника». Но цифры показывают, что правильный вопрос об инвестиции звучит не «работает ли чип?», а «кто купит часы его работы, по какой цене и сколько раз — и успеет ли это произойти до погашения капитальных обязательств?»

2. Новость, которая для разработчика означает скидку, для финансиста становится новой гипотезой

В сентябре 2026 года OpenAI объявила, что цены интерфейсов программирования приложений (Application Programming Interfaces; APIs) для GPT-6 Sol и GPT-6 Luna ниже сопоставимых промоцен предыдущего поколения, и частично связала это со снижением затрат на кэширование и инференс. В опубликованном примере цена

входных данных Sol снизилась с **4 до 2 долларов за миллион токенов**, а выходных — с **20 до 10 долларов**. Это сравнение двух продуктов разных поколений при заявленных условиях ценообразования, а не доказательство того, что рыночная арендная цена каждого чипа упала вдвое S04.

Создатель небольшого приложения увидит здесь пространство для эксперимента, которое прежде сужал счёт за инференс: можно проверить десять формулировок вместо пяти или обслужить пользователя, разговор с которым прежде не оправдывал более мощную модель. Тот, кто купил стойки процессоров и превратил их стоимость в долгосрочное финансовое обязательство, задаст другой вопрос: удвоится ли спрос вслед за падением цены услуги или придётся выполнять больше работы ради той же выручки?

Здесь возникает то, что я называю **ловушкой дефляции цены токена** (Token Deflation Trap). Само по себе снижение цены ещё не создаёт ловушку. Она появляется при совпадении трёх условий: выручка на проданную единицу работы падает; число полезных оплаченных единиц невозможно нарастить в нужной мере; стоимость привлечённого капитала не снижается столь же быстро. Стоит хотя бы одному условию не сработать — и ловушка может обернуться прекрасной возможностью. Поэтому название годится не как пророчество, а как вопрос, который удерживает в поле зрения главные сочленения задачи.

Рынок сложнее, чем «цена за миллион токенов». Входной токен отличается от выходного; токен, полученный из повторно используемого кэшированного контекста (Cached Input Token), не требует нового вычисления; удачный короткий ответ может заменить длинную цепочку попыток. Если продавать «миллион токенов», словно миллион одинаковых гвоздей, вскоре выяснится, что одни гвозди занимают память, другие требуют ожидания, а третьи вообще пришли из кэша. Учётная единица удобна; физическая единица непослушна.

3. Путь токена, которого не видно на экране чата

Чтобы понять, как программное обеспечение меняет цену оборудования, начнём с обычной пользовательской фразы: «Кратко изложи эту переписку и свяжи финансовое решение с письмами прошлой недели». Графический процессор (Graphics Processing Unit; GPU) получает не готовую к обдумыванию фразу, а последовательность токенов (Tokens), проходящих через вычисления, которые по-разному зависят от памяти, передачи данных и параллелизма.

На стадии первичного заполнения контекста (Prefill) входные токены обрабатываются так, чтобы построить представления механизма внимания (Attention). Затем начинается повторяющаяся генерация (Decode): выпускается один токен, после чего он включается в контекст следующего. В зависимости от размера нагрузки и архитектуры первая стадия обычно хорошо поддаётся параллельным вычислениям; вторая может упираться в пропускную способность памяти (Memory Bandwidth), перенос весов и состояние диалога. Чип, великолепный на первой стадии, не обязательно даст наименьшую итоговую стоимость на второй. Поэтому такие инструменты, как TensorRT-LLM, описывают архитектуры, где ресурсы двух стадий при необходимости разделены, а не принудительно работают в одном ритме [S07](#).

У диалога есть рабочая память, не похожая на долговременное хранение знаний: кэш ключей и значений (Key-Value Cache; KV Cache) удерживает представления, позволяющие не пересчитывать части контекста. Каждый длинный разговор требует дополнительного места. Если под него заранее отвести большой объём, а использовать лишь малую часть, внутри дорогого чипа будет копиться пустота, за которую никто не платит. Взгляд библиотекаря помогает понять: проблема не всегда в нехватке книг — иногда в том, как мы резервируем ящики.

В 2023 году исследователи представили систему vLLM и алгоритм страничного внимания (PagedAttention), вдохновлённый страничной организацией виртуальной памяти (Virtual Memory Paging) в операционных системах. Вместо расточительного непрерывного резервирования они разделили кэш внимания на блоки, которыми можно гибко управлять и делиться. В экспериментах рецензированной статьи пропускная способность инференса выросла примерно в два-четыре раза относительно отдельных базовых систем при сопоставимой задержке [S05](#). Это не означает, что любой центр обработки данных получит тот же результат после простой установки программы, но доказывает, что **алгоритмическое решение** способно перераспределить ценность уже имеющегося оборудования.

Ранее система Orca, представленная на конференции OSDI 2022, показала, что планирование на уровне одного цикла генерации (Iteration-Level Scheduling) может сократить ожидание коротких запросов за более длинными пакетами [S06](#). Непрерывная пакетная обработка (Continuous Batching), повторное использование префикса запроса (Prefix Caching) и разделение стадий вычисления — не косметические дополнения. Они меняют объём работы, который удаётся превратить в услугу за приемлемое время. Квантизация (Quantization) снижает точность числового представления и нагрузку на память и вычисления, но требует проверки качества: экономию пользователь может оплатить точностью ответа. При спекулятивном декодировании (Speculative Decoding) более быстрая функция предлагает токены, которые затем проверяет другая модель, поэтому коэффициент ускорения неодинаков для разных нагрузок.

Вот почему теоретический пик числа операций с плавающей точкой в секунду (Floating-Point Operations per Second; FLOPS) нельзя приравнивать к числу правильных ответов, которые клиент покупает за час. Между этими показателями стоят очередь, модель, планировщик, энергия, соглашение о времени ответа и запросы разной длины.

Эффективность — не постоянное свойство чипа, а поведение всей системы при выполнении конкретной работы.

4. Парадокс: программа может одновременно спасти инвестицию и снизить её цену

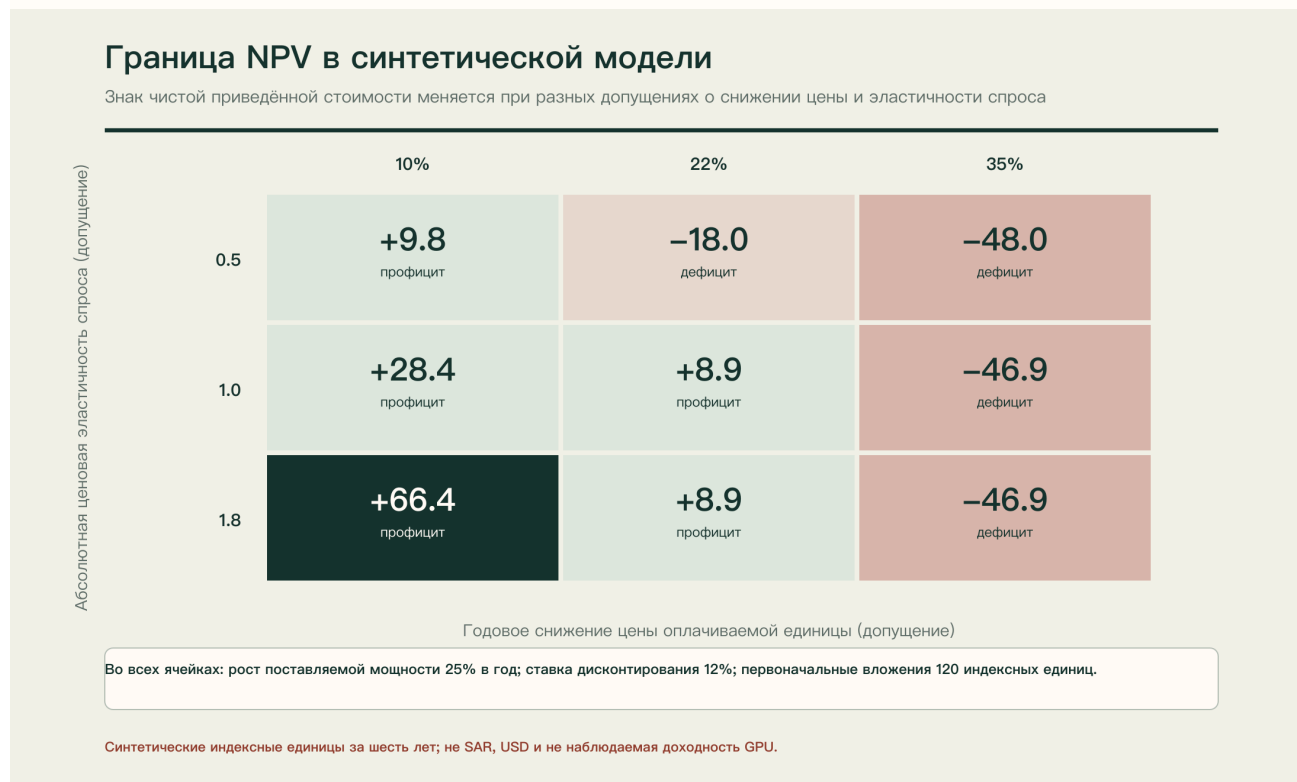
Возьмём условный пример, не относящийся ни к Саудовской Аравии, ни к какой-либо компании. Сервер обслуживал 50 тысяч единиц работы в день. После улучшения управления памятью его мощность без замены выросла до 80 тысяч. Прекрасный прирост производительности. Но если то же улучшение доступно всем конкурентам, вычисления становятся менее дефицитными, а цены переходят в новую фазу переговоров. Пользователь выигрывает дважды — от скорости и от конкуренции; владелец сервера вовсе не обязательно получает обе выгоды.

Вывод зависит от того, что экономисты называют ценовой эластичностью спроса (Price Elasticity of Demand): насколько растёт запрашиваемое количество при снижении цены? В упрощённой модели, если спрос пропорционален цене в отрицательной степени ϵ , то до учёта ограничений мощности и затрат выручка пропорциональна цене в степени $1-\epsilon$. При ϵ выше единицы общая выручка может вырасти вслед за падением цены; при ϵ ниже единицы она может сократиться, несмотря на рост потребления. А если мощность полностью загружена, весь расчётный энтузиазм спроса не имеет значения: дополнительные оплачиваемые единицы невозможно поставить.

Это точная версия вопроса, который Уильям Стэнли Джевонс (William Stanley Jevons) поставил в книге «Угольный вопрос» (The Coal Question) в 1865 году: может ли рост эффективности увеличить совокупное потребление ресурса вместо его сокращения? Он не предсказывал единицы искусственного интеллекта. И его историческое наблюдение не даёт права превращать его в лозунг «любое улучшение неизбежно удваивает спрос». **Эффект отдачи** (Rebound Effect) — возможность, зависящая от эластичности, доступной мощности, цены альтернативных приложений и ценности выполненной работы. Если вынести из этой истории одно правило, то такое: не стройте кривую спроса по силе восторга инженера перед процессором S08.

Поэтому рынку нужны два встречных вопроса, а не одно заявление. Откладывает ли программное обеспечение покупку следующего поколения, повышая доходность существующего оборудования? И делает ли оно услугу настолько изобильной, что

снижает цену продажи для всех? Оба ответа могут одновременно быть положительными, а победителя определяет не число токенов, вышедших из сервера, а денежная маржа, оставшаяся после их продажи.



Карта доходности в зависимости от эластичности и снижения цены; иллюстрационная симуляция, а не данные компании

Рисунок 1. Чистая приведённая стоимость (Net Present Value; NPV) в условных единицах, а не в долларах или риалах, для трёх темпов снижения цены и трёх значений эластичности спроса при предполагаемом ежегодном росте продаваемой мощности на 25%. Каждая ячейка получена запуском приложенной модели и не является прогнозом рынка.

5. Когда показатель загрузки мешает понять деньги

Инвестор иногда смотрит на загрузку графических процессоров (GPU Utilization), как хозяин гостиницы на занятые номера. Сравнение полезно, пока мы помним: гостиничный номер за ночь не превращается из люкса в стул в коридоре после обновления программы. И далеко не вся работа процессора подлежит оплате. Время может уходить на подготовку данных, повтор неудачного запроса, ожидание между сетью и памятью, внутреннюю работу, за которую клиент не платит, или результат недостаточного качества, который нельзя передать заказчику.

Процессор, загруженный на 95%, может приносить меньше выручки, чем процессор с загрузкой 65%, если различаются цена, ценность запросов, объём неоплачиваемых потерь или уровень обслуживания. И наоборот: после успешного программного улучшения загрузка может снизиться, потому что реальная мощность выросла быстрее спроса, — и падение показателя не означает операционный провал. Это вопрос знаменателя дроби, а не нравоучение о трудолюбии и лени.

В приложенной вычислительной модели условная мощность начинается с 50 единиц услуги в год, а цена единицы равна одному. Мы предполагаем капитальные затраты в 120 учётных единиц, горизонт в шесть лет, ставку дисконтирования 12% и явно заданные постоянные и переменные операционные затраты. Эти значения не взяты из расчётов какой-либо компании. Их достоинство в том, что они позволяют опровергать интуицию с лёгкостью, недоступной риторике.

При предполагаемом ежегодном падении цены на 22%, росте эффективной мощности на 25% и эластичности спроса 0,5 чистая приведённая стоимость оказывается **отрицательной примерно на 18 единиц**. При эластичности 1,0 она становится **положительной примерно на 8,9 единицы**. Повышение эластичности до 1,8 в этой паре допущений уже не улучшает результат: он остаётся около 8,9, поскольку оба значения спроса упираются в предел поставляемой мощности S12. Толпа заказов за дверью не оплачивает единицы, которые центр не может предоставить.

Мы также прогнали сетку из 27 сценариев: три темпа снижения цены, три значения эластичности и три скорости прироста производственной мощности. **В 12 случаях чистая стоимость была положительной, в 15 — отрицательной**. Эти числа не отражают вероятность прибыли: все ячейки — заранее выбранные синтетические условия проверки. Главный урок в том, что одинаково привлекательные суждения «дешёвый интеллект съест доходность чипов» и «рост спроса спасёт любые инвестиции» не выдерживают роли общих законов. Каждому требуется конкретное измеримое условие S12.

Есть и ещё один трюк, красивее трюка с загрузкой. В базовом сценарии моделирования потоки **номинально окупают капитальные затраты в пятом году**, но чистая приведённая стоимость остаётся **отрицательной примерно на 5,6 единицы**. Противоречия в расчёте нет: стоимость, вернувшаяся через пять лет, не равна стоимости, ушедшей сегодня, если капитал имеет цену. Проект может отпраздновать «окупаемость» и всё же при объявленной ставке дисконтирования уступить альтернативе. Фотография долгосрочного контракта не отвечает на вопрос о цене ожидания.

Нельзя прятать и наше оптимистичное допущение: переменная стоимость единицы снижалась тем же темпом, каким росла мощность. Когда в отдельном тесте чувствительности мы удержали её постоянной при прочих базовых допущениях, чистая стоимость упала с **–5,6 до –31,2 единицы**. Красивый слайд о «росте производительности» не заменяет измерение реального расхода энергии и операционной работы на услугу. Возможность быстрее выполнять больше работы не означает, что каждая работа обязательно подешевела \$12.



Динамика цены, объёма и выручки за шесть лет; условный пример

Рисунок 2. Три ряда из одного синтетического финансового эксперимента. Физическая динамика мощности расходится с коммерческой динамикой выручки; спрос показывает, сколько рынок хотел бы купить, а не сколько фактически поставлено. Полные границы графика и его параметры приведены в исследовательском приложении.

6. Саудовская Аравия: от проекта со стойками к экономике реальных клиентов

31 августа 2026 года AMD, Cisco и HUMAIN объявили, что вычислительная инфраструктура в Саудовской Аравии на процессорах AMD Instinct MI355X, сетях Cisco Silicon One и программном обеспечении ROCm **введена в эксплуатацию и обслуживает клиентов**. Те же стороны объявили план начать развёртывание мощности до **250 мегаватт** в 2027 году и ориентир расширения до **одного гигаватта** к 2030 году. Это три

разных состояния, и описывать их одной формулировкой нельзя: текущая эксплуатация по заявлению участников проекта, запланированное будущее расширение и дальняя цель, зависящая от исполнения, спроса и финансирования [S09](#).

Новость делает саудовский вопрос глубже. Появление процессоров AMD в потенциально платном сервисе вместо разговора, ограниченного закупкой NVIDIA, открывает путь к сравнению совокупной стоимости владения (Total Cost of Ownership; TCO), программной совместимости (Software Compatibility), стоимости переноса модели (Model Portability) и реальной производительности приложений на разных семействах оборудования. Но оно не доказывает, что одна платформа дешевле или прибыльнее другой, и не раскрывает клиентские контракты, загрузку или чистую маржу. Этих данных в объявлении нет.

Годом ранее HUMAIN объявила о запуске арабоязычного чат-приложения на модели ALLaM 34B с голосовым вводом на арабском, поиском и поддержкой арабского и английского языков. Из новости нельзя узнать ни число обслуженных токенов, ни выручку с одной сессии. Зато она переносит образ вычислительного центра из макета энергетической мощности в возможный сценарий на телефоне гражданина или в работе организации. Именно здесь происходит главный экономический переход: рынок вычислений существует не потому, что машины способны думать, а потому, что человек или учреждение нашли задачу, которую стоит им поручить и оплатить [S10](#).

Если Саудовская Аравия хочет перейти от покупки инфраструктуры к конкурентной позиции в индустрии инференса (Inference Industry), выбор между «самым крупным центром» и «самым дешёвым токеном» не должен завершать размышление. Предстоит найти виды высокоценной работы, которые можно выполнять с предсказуемым качеством и со стоимостью, снижающейся со временем. Это могут быть конвейеры, оптимизированные для арабского языка; специализированные модели с тестированием точности; услуги обработки с повышенной конфиденциальностью; разработка моделей с гарантированным временем ответа. Это примеры экономических категорий внутри вычислительной отрасли, а не заявление о том, что все они уже существуют или прибыльны.

7. Один тест может сделать пять похожих процессоров совершенно разными

Техническая реклама умеет сводить большие различия к одному слову — «производительность». Но инференс-движок, обслуживающий пять коротких вопросов одновременно, — не тот же движок, когда он ведёт тысячу длинных диалогов с повторным использованием контекста. Время до первого токена (Time to First Token; TTFT) отличается от задержки между токенами (Inter-Token Latency; ITL), а оба

показателя — от числа токенов в секунду (Throughput). Устройство может выиграть соревнование по токенам, медленно начать ответ и потому проиграть в приложении, где важна немедленная реакция. Другое уступит в скорости одиночного запроса, но победит при высоком параллелизме.

16 сентября 2026 года MLCommons выпустила обновление тестового набора MLPerf Inference v6.1. В него вошли новые испытания, более близкие к многошаговым задачам, а также результаты новых систем, проверяемые по правилам теста. Это важное пополнение знаний, но оно не выдаёт ни одному производителю справку о финансовой прибыльности. Тест измеряет рабочую производительность в заданных условиях, а не цену, которую оператор получает от клиентов, не стоимость его финансирования, налоги или расходы от простоя оборудования S11.

Представим реальное соревнование двух устройств одного поколения, где победитель заранее неизвестен:

- первое выдаёт четыре тысячи токенов в секунду на стандартной нагрузке, но его ежемесячная стоимость на клиента высока, потому что для гарантии быстрого первого токена приходится резервировать большой объём мощности;
- второе выдаёт лишь три с половиной тысячи токенов, зато совместно использует память контекстов, эффективно объединяет неоднородные запросы в пакеты и выигрывает от повторяющихся обращений пользователей, позволяя обслужить больше клиентов в пределах оплачиваемого ими времени.

Это **вымышленные иллюстрационные числа**, показывающие метод сравнения, а не измерения двух опубликованных устройств. Победителя нельзя назвать, пока мы не зафиксируем модель, длины входа и выхода, качество ответа, реальное распределение запросов, пределы задержки, стоимость электричества, охлаждения и эксплуатации, а также долю оплачиваемой работы. После этого «стоимость правильной принятой задачи» (Cost per Accepted Task) или «маржа услуги в рамках SLA» (Margin per SLA-Compliant Task) оказывается ближе к экономической жизни, чем сравнение устройств по изолированным вычислениям.

Именно здесь вопрос выходит из коробок с процессорами и попадает на стол директора по инвестициям: для какого набора нагрузок вообще подходит это устройство? Если клиентам нужна длинная последовательная генерация, ограничением становятся память и её пропускная способность. Если большая часть работы — поиск по общим корпоративным документам, особую ценность может дать повторное использование контекста. Если клиентам нужно обучение новых моделей, меняются распределение ценности и чувствительность к смене поколений. Неразумно складывать все эти счета в калькулятор под названием «цена часа», а затем удивляться разнице результатов.

8. Уловка, которой нельзя исправить договор на продажу электричества

В мире вычислений есть три скорости, которые трудно свести на один экран: скорость выпуска нового поколения, скорость подключения энергетической мощности и скорость обнаружения спроса, готового платить за результат. Первая движется циклами продуктов и исследований; вторая — строительством, сетями и поставками; третья — экспериментами организаций, которые могут начать с энтузиазмом, а затем обнаружить, что инференс нужен им не постоянно. Из этого разрыва возникает ценность **реального опциона** (Real Option): возможности отложить поставку, изменить тип нагрузки или временно арендовать мощность, вместо того чтобы превращать каждую будущую амбицию в сегодняшний счёт.

Обратимся к контракту, который не попадает на фотографии центров обработки данных. CoreWeave сообщает, что многие её многолетние договоры построены по принципу «бери или плати» (Take-or-Pay), а средневзвешенный срок твёрдых контрактов на конец 2025 года составлял около пяти лет. По описанным компанией условиям клиент платит за зарезервированную мощность, даже если использует её не полностью. Этот экономический механизм способен перенести часть риска снижения загрузки с оператора; поэтому ненаучно прогнозировать влияние снижения цен интерфейсов ИИ на его выручку так, будто все услуги немедленно продаются на открытом спотовом рынке [S01](#).

Но контракт — не волшебный щит и не игра, стирающая время. Его эффект зависит от способности компании выполнить обязательства, прав клиента, условий цены и продления, платёжеспособности сторон и графика долговых выплат. Если крупный клиент утратит экономическую потребность в зарезервированной мощности, он может платить до конца договора, а затем жёстко торговаться при продлении. Если вычислительный центр найдёт более выгодную возможность, он может оказаться не вправе перенаправить уже обещанную мощность. Маржа, которую приносит предсказуемость, при смене обстоятельств способна стать ценой упущенной возможности. Деталей саудовских контрактов у нас нет, поэтому автоматически переносить на них формулу американской компании нельзя.

Доходность процессора определяется не только суммой потоков, но и тем, когда возникают избытки и дефициты. Кредит, подлежащий выплате до созревания спроса, может создать нехватку денег даже при положительной чистой приведённой стоимости в будущем. Долгосрочный договор с фиксированной ценой защищает поток от падения цены, но не даёт в полной мере участвовать в её росте. Сильное программное улучшение может одновременно повысить маржу единицы и снизить выручку с неё. Эти обмены нельзя превращать в безликий список «рисков управления»: каждый

выражается рассчитываемой финансовой связью — графиком платежей, кривой спроса, пределом использования, датой нового технологического поколения и окном доступной электрической мощности.

9. Стратегическая идея: владеть не процессором как предметом, а кривой его экономического обучения

Возможно, в 2030 году главным активом вычислительного оператора окажется не то, что попадает в объектив экскурсии по центру обработки данных, а точная история, связывающая **тип нагрузки** с фактическими затратами, оплатой клиента и принятым им качеством. Это не декоративный журнал числа вызовов, а знание о том, где месяц за месяцем исчезало время и где возникала маржа.

Я называю это исследовательское предложение **«картой коммерческой жизни вычислений»** (Compute Commercial Life Map). Для каждого класса нагрузки в ней отмечаются пять разных часов: время адаптации модели к оборудованию, ожидание первого результата, выпуск приемлемого результата, срок конкурентной устойчивости цены и срок денежной окупаемости доли инвестиции. Эти часы не равны, и ни одни нельзя свести к общенациональному среднему. Физическая жизнь чипа может длиться шесть лет, ценовое преимущество — восемь месяцев, полезность на меньшей модели — значительно дольше, а срок загружающего его договора вообще не связан с первыми тремя.

Продавец машины обычно убеждает лучшим тестом. Опытному директору по инвестициям нужна **кривая угасания экономического преимущества** (Economic Advantage Decay Curve): сколько месяцев услуга остаётся ценнее альтернативы, прежде чем придётся снизить цену или сменить тип работы? Измерять это нужно по партиям оборудования, а не по названию поколения процессора. Тогда можно сравнить три действия, редко попадающие в заголовок гонки: улучшить программу, изменить цену услуги или найти клиентов, чьим задачам мощность подходит лучше. Возможно, отказ от дополнительной закупки окажется самым умным вычислительным решением года, хотя из него не получится эффектной первой фотографии.

Есть и заслуживающая уважения ставка против пессимизма: удешевление инференса способно породить приложения, которые прежде нельзя было продавать, увеличить рынок и повысить прибыльность несмотря на падение цены. Новые сложные задачи могут поглотить весь прирост, не дав начаться ценовой войне. Саудовской Аравии также могут помочь доступная энергия, инвестиционные возможности, региональные рынки и арабоязычные приложения. Но это условия и направления проверки, а не выводы из масштаба объявленного проекта. Пока мы не видим оплаченный спрос, маржу единицы и исполненные контракты, возможны обе истории.

10. Пять показателей, которые стоит прочитать до следующего объявления

Исследователь или руководитель по инвестициям уже завтра может начать с разрешённых агрегированных операционных и коммерческих данных, не ожидая несуществующего чудо-показателя:

Что измеряем	Какой вопрос это раскрывает	Ошибка при выборе самого простого числа
Выручка на принятую задачу (Revenue per Accepted Task)	Платит ли клиент больше полной стоимости успешного выполнения задачи?	Ограничиться ценой миллиона токенов при различии самих задач.
Эффективная пропускная способность при заданном уровне обслуживания (SLA-Constrained Throughput)	Сколько запросов можно выполнить с согласованными качеством и задержкой?	Использовать лабораторный максимум процессора.
Доля выручки по контрактам (Contracted Revenue Share)	Какая часть потока не зависит от спотовой цены и на каких условиях?	Приравнять объявленный контракт к полученным деньгам или гарантированной марже.
Срок окупаемости с учётом падения цены (Price-Adjusted Payback)	Возвращает ли проект капитал за время, пока сохраняется его коммерческое преимущество?	Ограничиться бухгалтерским сроком амортизации.
Распределение маржи по нагрузкам (Workload Margin Distribution)	Не скрывает ли среднее убыточные нагрузки и те, что их финансируют?	Показывать красивую общую загрузку, стирающую различия приложений.

Предположим открыто описанный тест на рынке Саудовской Аравии: два поставщика работают с сопоставимыми обезличенными нагрузками и соревнуются в предоставлении арабоязычной услуги известного качества в одинаковом временном пределе. Измеряется полная стоимость принятой задачи, затем опыт повторяется при изменении объёма спроса, распределения длины диалогов и доли повторно используемого контекста. Добавляется экономическая проверка: как снижение цены услуги на 20% повлияло на число привлечённых клиентов, завершённых запросов и прибыль? Раскрывать устройство критической инфраструктуры или клиентские тайны не требуется; достаточно опубликовать агрегированные сравнения и определения показателей. Исход заранее неизвестен: победить может более новое устройство, лучшая программа, более подходящая работа или сбалансированный контракт.

Гипотеза этой статьи такова: **«национальное превосходство в вычислениях искусственного интеллекта возникает не из владения наибольшей абстрактной мощностью, а из скорости превращения технических улучшений в полезную**

оплачиваемую работу до того, как снизится её цена». Это проверяемое утверждение. Если после учёта нагрузок, затрат и условий контрактов связи не обнаружится, его следует отбросить или сузить. Я не хочу строить рынок на красивой метафоре.

11. Что расскажет нам 2030 год, если сегодня мы правильно зададим вопрос?

В первом сценарии спрос на сложный инференс растёт быстрее, чем падает его цена; новое поколение оборудования открывает приложения, невозможные на предыдущем, а саудовские мощности получают выгоду от более крупного регионального рынка. Во втором программное обеспечение и новые компьютеры создают огромное изобилие, но большая часть ценности переходит к разработчикам приложений и пользователям; операторы вычислительных центров вступают в жёсткую конкуренцию по цене и затратам. В третьем оба сектора сосуществуют: сверхмощные системы обслуживают ограниченный круг дорогих задач, а универсальные вычисления становятся низкомаржинальным товаром. Это **условные траектории без вероятностных весов**, а не три рассчитанных прогноза.

Что наблюдать между 2026 и 2030 годами? Изменение фактических, а не только опубликованных цен; реальную долю твёрдых контрактов; состав входных и выходных токенов и кэша; производительность услуги при одинаковых временных ограничениях; разницу в стоимости задач на разных поколениях оборудования при той же нагрузке; долю выручки от клиентов за пределами внутреннего рынка. Если объём токенов растёт быстрее выручки, не следует сразу объявлять бедствие: сначала проверим маржу, затем окупаемость актива и стоимость привлечения клиентов. Если растёт выручка, не следует объявлять победу: сначала посмотрим, сколько из неё поглотили капитал, охлаждение и сети.

Директор старой фабрики слушал ритм машин и знал, что тишина может означать убыток. Директор современной фабрики интеллекта нуждается в иной смелости: увидеть, что процессор производит больше, а затем с крайней вежливостью и ещё большей строгостью спросить, кто оплатил этот выпуск, какая работа осталась неоплаченной и мог ли клиент выполнить тот же расчёт на меньшем устройстве.

Современная машина может выиграть каждый тест скорости и проиграть следующие переговоры о цене. Старое поколение может уступить в рейтинге производительности и оставаться прибыльным, потому что выполняет работу для хорошо известных клиентов. Противоречия здесь нет. Оно возникает у нас, когда мы смешиваем прогресс науки с прогрессом счёта.

И потому читателю стоит оставить в памяти вопрос не «сколько процессоров будет в Саудовской Аравии?», а «что Саудовская Аравия научится вычислять лучше других — и кому этот результат будет настолько нужен, что он заплатит за него даже после появления следующего процессора?»

Источники и упомянутые работы

1. [CoreWeave, 2025 Form 10-K, SEC, срок полезного использования и договорная информация, опубликовано в 2026 году.](#)
2. [CoreWeave, Form 10-Q, период до 30 июня 2026 года, SEC, Note 2 и RPO.](#)
3. [NVIDIA, объявление платформы Rubin, 5 января 2026 года.](#)
4. [OpenAI, Introducing GPT-6 Sol and Luna, сентябрь 2026 года, разделы о ценах и кэшировании.](#)
5. [Kwon et al., Efficient Memory Management for Large Language Model Serving with PagedAttention, SOSP 2023, DOI 10.1145/3600006.3613165.](#)
6. [Yu et al., Orca: A Distributed Serving System for Transformer-Based Generative Models, OSDI 2022.](#)
7. [NVIDIA, TensorRT-LLM, техническая документация Disaggregated Serving.](#)
8. [William Stanley Jevons, The Coal Question, 1865, оцифрованное историческое издание.](#)
9. [AMD, Cisco and HUMAIN, совместное объявление о вводе в эксплуатацию и расширении, 31 августа 2026 года.](#)
10. [Saudi Press Agency, HUMAIN Chat на основе ALLaM 34B, 25 августа 2025 года.](#)
11. [MLCommons, MLPerf Inference v6.1 Results, 16 сентября 2026 года.](#)
12. [Воспроизводимая оригинальная **синтетическая** модель, допущения и результаты 27 сценариев; см. \[таблицу результатов\]\(#\) и \[допущения\]\(#\).](#)

Редакционное примечание: опубликованные финансовые показатели приведены как факты с указанием источников; результаты моделирования не являются данными о компаниях или саудовских услугах. Упомянутые организации не участвовали в подготовке или рецензировании этой работы.