



СТРАТЕГИЧЕСКАЯ СТАТЬЯ

Редакционная дата: 23 октября 2026 года

Когда Даммам стал фабрикой секунд

Редакционная дата: 23 октября 2026 года

На демонстрации перед саудовской аудиторией примечательно было не то, что языковая модель ответила на вопрос о программе Vision 2030. На это способны десятки моделей. Джонатан Росс, основатель **Groq**, и Тарик Амин, возглавлявший тогда Aramco Digital, пытались показать на Всемирном саммите по искусственному интеллекту 2024 года другое: голосовой разговор с машиной может перестать быть обменом репликами, в котором человеку приходится сбиваться с ритма в ожидании ответа. Это была демонстрация двух компаний, а не опубликованное независимое испытание задержки. И всё же она точно обозначила вопрос, вокруг которого была построена целая компания: что меняется, когда пауза между двумя словами становится проблемой самой архитектуры процессора? [S01](#)

Несколько месяцев спустя появилась новость, не похожая на обычные сообщения о строительстве дата-центров. Groq заявила, что ввела в строй кластер инференса в Саудовской Аравии за **восемь дней в декабре 2024 года**. В феврале 2025 года компания сообщила, что её центр в Даммаме обрабатывает запросы клиентов со всего мира через **GroqCloud**; одновременно было объявлено о **саудовском инвестиционном обязательстве на 1,5 млрд долларов** для расширения инфраструктуры. Это сведения самой компании. Они не означают, что вся сумма уже потрачена, вся запланированная мощность построена или что опубликована бухгалтерская отчётность об экспорте услуг. Но и этого достаточно, чтобы изменить положение Королевства на технологической карте:

не каждый ответ, пришедший на саудовский телефон, обязательно вычислен за рубежом; запрос разработчика из другой страны, напротив, может быть обработан на саудовской территории. S02

Здесь возникает вопрос глубже торговли серверами. Почему место, где хранятся веса модели, порой менее важно, чем место, способное выдать следующий токен за долю секунды? Почему компания, создавшая специализированный чип для генерации речи, сделала Персидский залив узлом глобальной сети, а затем вступила с NVIDIA — своим, казалось бы, естественным конкурентом — в лицензионные отношения и архитектурное сближение? И есть вопрос, который нарушает стройность красивых цифр: если Королевство производит секунды меньшего ожидания, куда уходит затраченная электроэнергия и какая вода отводит образовавшееся тепло?

Эта статья не об отказах и не об управлении; «суверенитет модели» здесь не используется как лозунг.

Эта статья посвящена **производству времени инференса** (*inference*). Оно начинается в электрической тишине небольшой памяти внутри чипа, проходит через компилятор (*compiler*), планировщик и тепловые контуры и заканчивается счётом за ресурсы, который можно проверить. В центре — неочевидная возможность для Саудовской Аравии: **экспортировать измеримую временную производительность, а не просто размещать файл под названием «модель»**.

Ответ не возникает целиком

Представим, что вы попросили голосового помощника рассказать о промышленном отчёте. Слушатель слышит цепочку предложений; система инференса видит последовательность **токенов** (*tokens*), которые вероятностно выбирает нейронная сеть. Сначала запрос проходит стадию **предварительной обработки контекста** (*prefill*): токены контекста обрабатываются относительно параллельно, создаются внутренние представления и кэш ключей и значений внимания (*key-value cache, KV cache*). Затем начинается **авторегрессионная генерация** (*decode*): модель выдаёт один токен, и тот становится частью контекста, необходимого для следующего. Это не литературное деление времени, а две разные физические задачи вычисления.

На стадии *prefill* преобладают умножения матриц и крупные вычислительные операции. *Decode* одного запроса требует многократного обращения к весам модели и состояниям внимания; его может ограничивать перемещение данных через память сильнее, чем собственно арифметика. Поэтому заявления о «большом числе триллионов операций в секунду» недостаточно. Точный вопрос звучит так: сколько вычислительной работы выполняется на каждый байт, полученный из памяти?

Инженеры называют это **арифметической интенсивностью** (*arithmetic intensity*) и используют **модель Roofline** (*Roofline model*), чтобы понять, когда производительность упирается в вычислительные блоки, а когда — в пропускную способность памяти.

Производительность завода можно поднять, добавив линии, но это ничего не говорит клиенту, который ждёт свой заказ сейчас. Так же различаются **совокупная пропускная способность** (*aggregate throughput*) и **скорость генерации для одного пользователя** (*tokens per second per user*). Поставщика может радовать огромное число, полученное после объединения тысяч запросов, пока отдельному пользователю кажется, что модель говорит с ним неторопливо. Между этими двумя впечатлениями лежат четыре разных показателя: время до первого токена (*time to first token, TTFT*), межтокенная задержка (*inter-token latency*), ожидание в очереди и полное время от отправки запроса до последнего токена (*end-to-end latency*). В эксплуатационной документации Groq эти показатели разделены; различаются также услуги с повышенными требованиями к задержке и пакетные задания, которым мгновенная реакция не нужна. [S09](#) [S10](#)

Возьмём расчётный пример, не связанный со скоростью какого-либо конкретного чипа. Если нужно сгенерировать 180 токенов со скоростью 300 токенов в секунду, теоретическое время генерации составит 0,6 секунды. При 900 токенах в секунду — 0,2 секунды: выигрыш только на этой стадии равен 0,4 секунды. В расчёт не входят передача речи модели, распознавание голоса, обращения к внешним инструментам и ожидание до начала генерации. Но если программный агент выполняет четыре последовательных цикла генерации того же размера, накопленная экономия этой стадии может достичь 1,6 секунды. В **агентных системах** (*agentic systems*) эффект скорости последовательной генерации умножается на число шагов. Поэтому экономика инференса отличается от экономики однократного обучения модели.

Зачем Groq перенесла расписание внутрь чипа

Чип, на котором компания построила свою концепцию, называется **процессором языковой обработки** (*Language Processing Unit, LPU*). Название скорее служит коммерческим обозначением специализированного класса вычислителей, чем общепринятым архитектурным термином, но стоящая за ним инженерия заслуживает разбора. Современные GPU мощны и универсальны. Они используют память с высокой пропускной способностью (*high-bandwidth memory, HBM*), Tensor Cores, планировщики, наборы инструкций и программные слои для широкого круга задач глубокого обучения. Groq, напротив, стремится сделать **порядок перемещения данных и операций в значительной степени известным до начала исполнения**. [S04](#) [S05](#)

В традиционной архитектуре программа во время исполнения выясняет, какие ресурсы готовы выполнить работу. При **статически спланированном потоке данных** (*statically scheduled dataflow*) компилятор размещает операции и перемещения данных на временной шкале. Результаты проходят через вычислители, память и соединения, как по производственной линии, где заранее известно время прибытия материала к каждой станции. Это толкование заявленной архитектуры, а не обещание детерминизма всей внешней услуги: очередь, сеть, конкуренция пользователей и различия между моделями по-прежнему создают разброс. Предсказуемое движение данных внутри машины лишь снимает часть решений, которые иначе пришлось бы принимать аппаратуре во время исполнения. [S04](#)

Компания выбрала **статическую память с произвольным доступом на кристалле** (*Static Random-Access Memory, SRAM*) вместо преимущественной загрузки весов из HBM вне вычислительного кристалла. SRAM хранит биты в быстрых транзисторных ячейках и не требует циклов обновления, подобных DRAM, но каждый её бит занимает больше площади кремния. Низкая задержка и скорость покупаются ценой плотности хранения. В этом компромиссе и состоит смысл архитектуры, а не в одном лозунге о «памяти в десять раз быстрее». Для одного поколения чипов Groq указывает пропускную способность SRAM свыше 80 ТБ/с и сопоставляет её примерно с 8 ТБ/с у некоторых конфигураций GPU/HBM. Это **маркетинговое сравнение, зависящее от поколения и определения тракта данных**; оно не позволяет судить обо всех GPU и не доказывает десятикратного ускорения полного инференса любой модели. [S04](#)

Самая трудная задача звучит просто: **куда поместить веса?** Гипотетической модели с 70 млрд параметров при 16-битном представлении нужно около 140 млрд байт только под веса, то есть **140 десятичных гигабайт** — без состояний внимания, временных областей, репликации и межчиповой координации. При восьми битах сырой объём снижается до 70 ГБ, но развёртывание не становится бесплатным: остаются численная точность, масштабы **квантования** (*quantization*), KV cache и способы распределения. В 2026 году NVIDIA объявила о стойке **Groq 3 LPX** с 256 ускорителями, **128 ГБ SRAM суммарно** и заявленной внутристоечной пропускной способностью памяти 40 ПБ/с. Даже в этом новом поколении необработанные 16-битные веса модели 70B могут не уместиться в SRAM одной стойки. Это не означает, что система не способна обслуживать модель: существуют разбиение весов, квантование, другая память и другие компоненты. [S08](#)

Вывод: чем ближе данные к вычислению, тем короче их доставка, но тем важнее рассматривать программу как **карту размещения весов, памяти и межчиповых коммуникаций**. Компетенция, которую можно развивать локально, начинается не с покупки чипа, а с размещения с учётом компилятора (*compiler-aware placement*), квантования без неприемлемой потери качества, планирования пакетов переменной

длины и измерения частоты перемещений KV-данных между ресурсами. Чтобы создать сильную команду в этих эксплуатационных слоях, не нужно за один день изобретать собственный чип.

Концептуальное сравнение потока памяти в GPU и LPU

Рисунок 1. Авторская пояснительная схема; она не является производственной диаграммой и не измеряет производительность двух архитектур. Источники принципов: [S04](#), [S07](#), [S08](#).

Три реальных опыта и три разных вида «успеха»

Первый опыт: от сцены к Даммаму. В сентябре 2024 года Aramco Digital и Groq объявили о сотрудничестве по созданию крупного центра инференса и намерении связать его с рынком цифровых услуг **nawat**. На сцене показали голосовой диалог. Это подтверждение работы узкого демонстрационного сценария, а не независимое свидетельство существования всей запланированной инфраструктуры. Затем появились сообщения о декабрьском развёртывании кластера за восемь дней и о предоставлении GroqCloud из центра в Даммаме в феврале 2025 года. Различие существенно: соглашение создаёт возможность, демонстрация подтверждает ограниченный сценарий, а объявленная облачная услуга указывает на эксплуатацию продукта для разработчиков. Доступные публичные документы не раскрывают почасовое потребление электричества и воды или число токенов, произведённых только в Даммаме. [S01](#) [S02](#)

Второй опыт: локальной модели не нужен локально произведённый чип, чтобы работать здесь. На LEAP 2025 Groq показывала работу саудовской модели **ALLaM**, моделей рассуждения и генерации речи на английском и арабском языках. В мае, согласно заявлению компании, она стала официальным поставщиком инференса для саудовской **HUMAIN**. Первое событие показало модели и сценарии, второе обозначило отношения поставщика. Здесь важно не смешивать три слоя цепочки создания стоимости: **модель может быть разработана в Саудовской Аравии, процессор — спроектирован и произведён в США**, как сообщает Groq, **а местом вычисления запроса может быть Саудовская Аравия**. Национальная рамка не превращает эти слои в один. [S02](#) [S03](#)

Третий опыт: конкурент стал частью одной машины. В декабре 2025 года Groq объявила о неисключительной лицензии NVIDIA на свою технологию инференса. Джонатан Росс, Санни Мадра и инженеры перешли в NVIDIA, а Groq, по собственному заявлению, осталась независимой компанией. В августе 2026 года NVIDIA сообщила, что **Groq 3 LPX** запущена в серийное производство и что первым облачным провайдером, внедрившим

её, стала **Nebius**. Groq, в свою очередь, заявила, что **планирует** войти в число первых пользователей в сотрудничестве с Dell. «Продукт запущен в производство» не означает «установлен в саудовском центре»: опубликованных доказательств этого нет. [S06](#) [S07](#) [S12](#)

Новая архитектура NVIDIA объединила две семьи вычислителей: GPU-системы **Vera Rubin NVL72** и специализированные стойки LPX. По описанию NVIDIA, обе участвуют в вычислении слоёв модели, обслуживая длинный контекст и быструю генерацию. Это свидетельствует о том, что **инференс не обязательно представляет собой одну операцию для одного типа кремния**. Плотную обработку контекста можно поручить одной вычислительной архитектуре, а последовательную генерацию токенов — другой, распределив работу, сеть и память на более мелком уровне. NVIDIA со ссылкой на Artificial Analysis опубликовала результат около **3400 выходных токенов в секунду** для *Gemma 4 31B* при контексте 100 тыс. токенов. Он относится к этой модели, испытанию и платформе, а не обещает такую скорость каждому пользователю и не является результатом из Даммама. [S07](#) [S08](#)

География добавляет ещё одну оговорку. Сообщения 2025 года описывают работу из **Даммама**, тогда как современная страница платформы Groq показывает точку **SA-1 — Riyadh** и заявляет о 13 глобальных центрах. Из страницы непонятно, идёт ли речь о другой эксплуатационной метке, новом объекте или изменении классификации точек доступа. **Мы не считаем Даммам и Эр-Рияд двумя отдельными центрами и не предполагаем перенос из одного города в другой**. Достоверно лишь то, что Саудовская Аравия стала узлом заявленной сети Groq; точная конфигурация и текущее распределение нагрузки неизвестны. [S02](#) [S08](#)

Не каждый токен, вышедший из чипа, пришёл вовремя

Если процессор выдаёт тысячу токенов в секунду, но долго ждёт допуска из очереди, приложение теряет время, которого нет в спецификации чипа. Если агент вызывает четыре инструмента на разных континентах, ускорение decode не убирает сетевую задержку и время работы инструментов. Поэтому **карта времени запроса** полезнее одного рекордного показателя: круговая задержка сети (*round-trip time*), очередь, prefill, интервалы decode, ожидание инструментов и доля полностью выполненных задач при заданном качестве.

Коммерческая документация Groq показывает практическое различие временной ценности работы: уровень **performance** предназначен для строгих требований к отклику, **on_demand** — для обычного спроса, **flex** меняет часть гарантий обслуживания на более широкие пределы обработки; отдельно существует **Batch API**. Это не доказывает, что в компании уже работает тепловой планировщик, но показывает, что

реальная услуга инференса различает работу по срочности. Ответ в телефонном разговоре нельзя отложить до более прохладной погоды. Сводку большого архива или серию оценок модели можно перенести на несколько часов, если это допускают договор и характер данных. [S09](#) [S10](#)

Отсюда можно определить новый товар саудовского облачного оператора: **токен приемлемого качества, доставленный в установленный срок, с учётом энергии и воды, затраченных на его получение**. Сырое число токенов вводит в заблуждение: система способна генерировать ненужные длинные ответы, повторять неудачный результат или завершать пакет после срока. Энергия простаивающего чипа и охлаждения вообще не отражается в счётчике токенов. Нужны три пересекающихся измерения: полезная работа, соблюденный срок и материальный ресурс.

Погода — не фон здания, а возможный вход алгоритма

В информатике размещение задачи в подходящем временном окне называется **планированием** (*scheduling*). Что изменится, если включить наружный воздух в функцию стоимости? Температуры на экране телефона недостаточно. Инженер различает **температуру сухого термометра** (*dry-bulb temperature*) и **температуру влажного термометра** (*wet-bulb temperature*), связанную с возможностью испарения при данной влажности. Горячий сухой воздух допускает иное испарительное охлаждение, чем влажный прибрежный; прибрежная ночь необязательно «холодна» с точки зрения способности воздуха принять дополнительный пар.

Электрическая энергия в центре инференса не исчезает. Почти вся она в итоге превращается в тепло, которое нужно отвести от процессоров, памяти и блоков питания. Тепло можно передавать жидкости через **холодные пластины** (*cold plates*), затем через теплообменники и насосы — внешней системе отвода. Замкнутый жидкостный контур **не означает нулевого расхода воды**, как и жидкостное охлаждение не обязательно означает использование испарительных башен. Водный след зависит от способа отвода тепла, подпиточной воды, обработки и сброса, а также от источника электричества. [S11](#) [S13](#)

Сухое охлаждение отводит тепло в воздух с небольшим прямым испарением воды на площадке, но в жару может требовать больше электричества. **Испарительное охлаждение** в некоторых условиях снижает электрическую нагрузку ценой расхода воды на месте. Универсально выигрышного режима нет. Министерство энергетики США определяет **PUE** как отношение всей энергии объекта к энергии ИТ-оборудования, а площадочный **WUE** — как количество воды объекта на киловатт-час энергии ИТ. Ни один показатель сам по себе не сообщает, сколько полезных задач выполнено и сколько воды ушло на производство электричества за пределами объекта. [S11](#)

Физику помогает понять небольшое наблюдение. Для испарения килограмма воды примерно при 25 °C требуется около **2,44 МДж** скрытой теплоты; киловатт-час равен 3,6 МДж. В **идеальном абстрактном** случае, когда всё тепло одного киловатт-часа отводится испарением, теоретически испарится около **1,47 литра воды**. Это не WUE конкретного центра, не расход воды на один вопрос и не учёт продувки, потерь или электрической цепочки. Это связь энергии с фазовым переходом вещества, напоминающая: биты невесомы, но их тепло — нет. [S14](#)

Воспроизводимый опыт: может ли планировщик перенести тепло на другой час

Для статьи выполнен небольшой расчётный опыт: **исполнение реальное, все входные данные синтетические**. Мы построили окно из 48 часов и два вымышленных погодных ряда — «жаркий сухой» и «влажный прибрежный». Они не названы Эр-Риядом и Даммамом, потому что **почасовых саудовских наблюдений в модели нет**. Это математически правдоподобные профили, а не метеоданные. В каждый час включена интерактивная нагрузка, которую нельзя перенести, и пакетное задание с допустимой задержкой до девяти часов. Заданы предел мощности ИТ, постоянное потребление в простое и функции электричества и воды для сухого и испарительного охлаждения. Расчёт двухэтапный и разделимый: для каждого часа сначала независимо выбирается режим охлаждения с минимальной линейной стоимостью, затем **линейная программа** (*linear programming*) распределяет пакетную работу по часам, сохраняя сроки и **одинаковый синтетический объём работы**. Такая схема равносильна совместному выбору только при принятых здесь линейных разделимых затратах, бесплатном переключении режимов и отсутствии ограничений на минимальное время их работы. Коэффициент обмена литра воды на киловатт-час — математический вес, а не саудовская рыночная цена. Все допущения и уравнения приведены в исследовании и сопровождающем коде. [S17](#)

В жарком сухом сценарии политика, запускающая пакеты при первой возможности и выбирающая наименее энергоёмкое охлаждение, потребила за 48 часов **5192,71 кВт·ч и 4829,34 л воды**. Двухэтапная оптимизация по электричеству снизила расход до **5179,88 кВт·ч** и примерно **4167 л**. При весе воды **0,02 кВт·ч/л** решение перешло к **5188,14 кВт·ч и 3559,52 л**. По сравнению с минимумом электричества около **607 л** экономятся ценой примерно **8,26 кВт·ч**. Это не реальная норма экономии и не обобщаемый процент: результат зависит от придуманных для исследования функций охлаждения. [S17](#)

Во влажном прибрежном сценарии компромисс иной. Решение с минимумом энергии использовало **5245,08 кВт·ч и 4893,67 л**; с тем же весом воды — **5263,52 кВт·ч и 3640,80 л**. В этом синтетическом мире примерно **18,43 кВт·ч** добавляются, чтобы сократить расход примерно на **1253 л**. Все 48 пакетных заданий в обоих сценариях завершены в срок.

Кроме того, каждая политика обрабатывает одинаковый синтетический объём $W \approx 4394,2$ условной единицы: интерактивная и пакетная нагрузка складываются на общей искусственной шкале с равным весом единицы. Это не число задач и не прямое значение показателя принятых задач Y , определённого в исследовании. Числа показывают возможный механизм: существенна не формула «холоднее — лучше», а **согласование времени выполнения с предельной стоимостью тепла и воды**. S17

Граница компромисса между водой и электричеством охлаждения в двух синтетических сценариях

Рисунок 2. Результаты фактического решения линейной программы на вымышленных входных данных. Каждая точка обслуживает один и тот же объём работы без нарушения сроков. Оси относятся к синтетической модели, а не к саудовскому объекту.

Интерактивная работа в опыте вообще не перемещается, поэтому компромисс не достигается замедлением живого разговора. Двигается лишь работа, у которой действительно есть временное окно. Тепловой планировщик — не общий переключатель для всех запросов: его возможности определяются **доступной временной гибкостью**. Если гибкости нет, программа не должна делать вид, что способна изменить погоду.

Перенос пакетной работы в течение синтетических суток

Рисунок 3. Пример времени пакетной работы в жарком сухом сценарии. Линии получены локальным расчётом и не представляют реальные нагрузки Groq или HUMAIN.

Между планировщиком компилятора и планировщиком города

Работа Groq внутри чипа и предлагаемый внешний подход связаны логикой координации, но выполняют разные функции. Компилятор размещает результаты матричных операций по узлам чипа; он не может ждать, пока остынет город, прежде чем выдать следующий токен живому пользователю. Планировщик услуги решает, в какой час, на какой площадке и при каком охлаждении выполнять пакет оценки или обработки документов. Первый действует в наносекундах и тактах, второй — в минутах и часах. Их общий вопрос: **какую работу можно переместить, не нарушив условия её правильности и срока?**

Инженерная честность требует признать предшествующие работы. Исследование **Balancing Bits and Drops**, опубликованное как препринт в 2026 году, уже рассматривает планирование нагрузки с учётом воды и локального дефицита. В августе исследователи

представили **PowerSlider** для ограничений мгновенной электрической мощности в SGLang: снижение частоты GPU неодинаково влияет на prefill и decode; авторы опубликовали экспериментальные показатели goodput для своей установки, не для саудовского центра и не для LPU. Предлагаемая проверка не изобретает «экологическое планирование» заново. Её предмет уже: **связать меняющееся во времени охлаждение с реальными характеристиками нагрузки инференса на саудовских объектах и оценивать работу, выполненную в срок и с нужным качеством, сопоставимыми измерениями.** [S15](#) [S16](#)

Эксплуатационная команда начала бы со следа запроса: объёма prefill, числа выходных токенов, размера KV cache, числа инструментальных шагов, времени начала, первого токена и крайнего допустимого завершения, количества повторов. Физический слой добавил бы мощность PDU стойки, состояния насосов и вентиляторов, эффективность теплообменника, температуры сухого и влажного термометра, счётчик подпиточной воды и пределы расхода охлаждающей жидкости. Возможно, точное квантование или **кэширование общих префиксов** (*prefix caching*) окажется важнее переноса задания на вечер. Возможно, разделение prefill и decode по специализированным ресурсам важнее всех тепловых мер. Приоритет устанавливают измерения полезных токенов на единицу ресурса, а не риторика.

Где может ошибаться сама идея

Перенос пакетов способен скопить нагрузку, поднять пик или снизить использование серверов; если чипы остаются включёнными в простое, ожидаемой экономии не будет. Перемещение состояния KV или повторная загрузка модели в память менее занятого ускорителя могут потребовать больше энергии и времени, чем сэкономят охлаждение. Производительность системы охлаждения может зависеть от температуры подаваемой воды, а не только воздуха. Режим, лучший по площадочной энергии, может оказаться хуже по воде, связанной с внешним производством электричества; местного WUE тогда недостаточно.

Есть и более сильное экономическое возражение. Если Groq и её аналоги ускоряют инференс, приложения могут увеличить потребление токенов вместо снижения общего счёта. Быстрый агент позволяет пользователю запускать больше циклов чтения, эксперимента и проверки. **Энергия на токен** способна снижаться, пока **общая энергия** растёт или меняется сам состав работы. Обзор Лаборатории Лоуренса в Беркли 2025 года показывает: при разных предпосылках и объектах водопотребление, приписываемое нагрузке, может различаться более чем в десять тысяч раз из-за эффективности серверов, водоёмкости электросети, загрузки, охлаждения и климата. Это не измерение для Саудовской Аравии, а научное предостережение против вымышленной постоянной «литров на вопрос». [S13](#)

Для живого голосового ответа пользователю на другом континенте близость сети может быть важнее экономии охлаждения на далёкой площадке. Возможность глобально экспортировать вычисления ещё не делает место лучшим для каждого запроса. Одни приложения чувствительны к сетевой задержке, другие допускают пакетную отправку и последующее получение результата. Такая отрасль нуждается в двух коммерчески разных нагрузках: мгновенной, покупающей доли секунды, и гибкой, покупающей более дешёвый ресурс в пределах срока. Первую нельзя обременять всеми водными экспериментами, которые вторая может пройти без ущерба для пользователя.

Что можно построить в Саудовской Аравии в 2027–2035 годах

До присвоения громких титулов можно провести один промышленный опыт в саудовском центре инференса. Следует выбрать три реальные нагрузки с ясными соглашениями об уровне обслуживания: голосовой разговор в реальном времени, многошаговый программный помощник и пакетную оценку модели со сроком восемь часов. Каждая запускается на GPU и LPU, если это допускают модели, лицензии и доступность, при **одинаковом качестве, числе действительных токенов и распределении длины запросов**. Измеряются p95 TTFT, распределение межтокенной задержки, полезная работа в час, электричество стойки и объекта, площадочная вода и, при наличии достоверных данных, вода для производства электроэнергии. Опыт повторяется в разные сезоны, а не в один демонстрационный день.

Здесь нет призыва за три года с нуля создать саудовский LPU. Groq прямо заявляла о разработке, производстве и сборке своих чипов в Северной Америке. Более ценный локальный вопрос: могут ли саудовские команды стать сильными в **компиляции инференса для разнородного оборудования** (*heterogeneous inference compilation*), распределении контекста и KV cache и испытаниях, которые делают энергию и воду **на одну правильную задачу, выполненную в срок**, проверяемыми? Можно ли создать ПО, выбирающее площадку и платформу по ожидаемой длине ответа, сетевой стоимости передачи токенов и погоде, а затем экспортировать эту компетенцию вместе с услугой инференса? Это задача исследований, разработки и эксплуатации, а не уже полученный результат.

Практическая отправная точка скромнее: синхронная система измерений и А/В-опыт в одном центре, где небольшая доля гибкой нагрузки передаётся **планировщику, учитывающему погоду**. Если он снижает расход воды без ухудшения отклика и без переноса нагрузки на более водоёмкое электричество, получен реальный выигрыш. Если нет, отрицательный результат тоже полезен: он указывает решающее ограничение системы. Успех должен выражаться ресурсами и завершённой работой, а не красивым движением линии на графике.

В прошлом веке некоторые отрасли экспортировали энергию в форме алюминия, металлов и удобрений, преобразуя материю с большим вкладом электричества. В следующие десятилетия может возникнуть отрасль, превращающая электричество, кремний и охлаждение в **оплачиваемое когнитивное время**: миллисекунды, снятые с голосового диалога, или тысячи попыток вывода, которые научный агент успел закончить к сроку. Это **предложенная экономическая аналогия**, а не уравнивание отраслей. В обоих случаях конкурентоспособность определяется тем, насколько хорошо ресурсы преобразуются в продаваемый результат; потери тепла и ожидание памяти становятся дефектом товара, а не числом в аппаратной комнате.

Так мы возвращаемся к первой сцене. Groq показала разговор, который казался быстрее обычного. Но оставшийся после него стратегический вопрос касается не одной скорости искусственного голоса: **что, если Саудовская Аравия сможет производить для мира меньше времени ожидания и для каждой задачи знать электрическую и водную цену этого времени?** Тогда дата-центр станет инженерным производством с определённым выпуском, чип — одним из его инструментов, а планировщик будет читать погоду наряду с длиной фразы. Ответ не даст фотография новой стойки. Его дадут эксперименты, связывающие токены с завершёнными циклами, циклы — со счётчиками, а измерения — с результатом, который другой инженер способен пересчитать.

Редакционное примечание и источники

Это самостоятельная статья. Сведения Groq и NVIDIA о производительности и расширении отнесены к заявлениям компаний, если независимая проверка не указана; географическая неопределённость раскрыта. Все числа 48-часового опыта получены из воспроизводимого синтетического расчёта в сопровождающем архиве и не являются конфиденциальными эксплуатационными данными или измерениями в Саудовской Аравии.

Основные источники

1. S01 — Groq и Aramco Digital: объявление о партнёрстве, 12 сентября 2024 года.
2. S02 — Groq: Даммам и обязательство на 1,5 млрд долларов, 10 февраля 2025 года.
3. S03 — Groq: поставщик инференса для HUMAIN, 27 мая 2025 года.
4. S04 — Groq: устройство LPU, 7 марта 2025 года.
5. S06 — Groq и NVIDIA: неисключительная технологическая лицензия, 24 декабря 2025 года.
6. S07 — NVIDIA: запуск Groq 3 LPX в серийное производство, 24 августа 2026 года.
7. S08 — платформа и сеть Groq.

8. S09 — уровни обслуживания Groq и пакетный интерфейс.
9. S11 — Министерство энергетики США: эффективность охлаждения, PUE и WUE.
10. S13 — Лаборатория Лоуренса в Беркли: водопотребление нагрузок дата-центров, 2025 год.
11. S15 — Talukder et al., *Balancing Bits and Drops*, препринт 2026 года.
12. S16 — Li et al., *PowerSlider*, препринт 2026 года.