



ИССЛЕДОВАТЕЛЬСКАЯ СТАТЬЯ

Редакционная дата: 23 октября 2026 года

От фабрики токенов к тепловому планировщику: операционное исследование экономики инференса в Саудовской Аравии

Редакционная дата: 23 октября 2026 года

Концептуальная и вычислительная рабочая статья

Жанр исследования. Концептуальная и вычислительная рабочая статья (*conceptual and computational working paper*) с фактически выполненным линейным программированием на вымышленных данных, обзором первичных источников и заявленных эксплуатационных событий. Это не полевое исследование объектов Groq, не независимый бенчмарк её чипов и не национальная оценка потребления воды.

Аннотация. В статье рассматривается возможность перейти в управлении языковым инференсом (*inference*) в Саудовской Аравии от решения о числе ускорителей к совместной оптимизации полезной вычислительной работы, времени, электричества и воды для охлаждения. Эмпирическая отправная точка — объявленные Groq сотрудничество с Aramco Digital в 2024 году и работа GroqCloud из Даммама в 2025 году, а техническая — **процессор языковой обработки** (*Language Processing Unit, LPU*) со **статической памятью с произвольным доступом на кристалле** (*Static Random-Access Memory, SRAM*) и статическим планированием перемещения данных. В дальнейшем последовали неисключительное лицензирование технологии NVIDIA в декабре 2025 года и появление платформы Groq 3 LPX в августе 2026 года. Эксплуатационные

события отделены от планов расширения и результатов, сообщённых поставщиками. Модель различает предварительную обработку контекста (*prefill*) и последовательную генерацию (*decode*), мгновенные запросы и пакетную работу со сроками, воду на площадке и воду, связанную с электросетью. Выполнен 48-часовой двухэтапный разделимый расчёт для двух синтетических погодных режимов: для каждого часа режим охлаждения сначала выбирается независимым минимумом линейной стоимости, затем линейная программа распределяет гибкую работу. Равносильность совместному выбору действует только при принятых здесь линейных разделимых затратах, бесплатном переключении и отсутствии межчасовых ограничений режимов. Все решения обработали приблизительно **4394,2 условной единицы синтетического объёма работы W** на общей искусственной шкале без нарушения сроков; это не число задач Y . В вымышленном жарком сухом режиме увеличение веса воды в целевой функции снизило её расход примерно на 607 л ценой около 8,26 дополнительных кВт·ч относительно решения с минимумом энергии. Это процедурное доказательство возможности построить многоресурсный оптимизатор, а не измеренный эффект объекта и не доказательство превосходства LPU. В заключение предложен опровержимый полевой опыт, допускающий победу конкурирующей архитектуры и не предполагающий заранее, что тепловое планирование полезно.

Ключевые слова: Groq; Language Processing Unit (LPU); статический поток данных; SRAM; HBM; prefill/decode; KV cache; температура влажного термометра; PUE; WUE; водопотребление; инференс; линейное программирование; планирование пакетной работы; Саудовская Аравия.

1. Проверяемый вопрос и границы утверждения

Центральный вопрос исследования: **если саудовский оператор инференса обслуживает как жёстко ограниченный по времени интерактивный спрос, так и отложенную работу, а также располагает достоверными данными о влиянии нагрузки на электричество, способы охлаждения и погоду, способен ли он снизить расход воды или электроэнергии на одну правильно выполненную задачу, не нарушив требований ко времени и качеству?** Научно обоснованный ответ нельзя считать положительным до получения данных от реального оборудования и нагрузки. Поэтому здесь предложены три результата с более узкими притязаниями и большей проверяемостью:

1. Реконструкция саудовского кейса Groq по её объявлениям с классификацией доказательств и различием наблюдаемого события, соглашения, дорожной карты и теста поставщика.
2. Причинно-техническая модель, связывающая длину входа, генерацию, память, планировщик, энергию серверов, отвод тепла, погоду и расход воды.

3. Исполненный вычислительный опыт на вымышленных данных, показывающий пространство компромиссов и пределы обобщения, а затем протокол замены придуманных функций измерениями реального объекта.

Мы не предполагаем, что **более дешёвый инференс на токен** неизбежно экономнее на задачу или что меньшая энергия всегда означает меньше воды. Специализированный чип не считается лучшим для каждой нейросети по определению. 48-часовой опыт не приписывается Groq или HUMAIN: он создан для этой работы и не воспроизводит известную инфраструктуру любой из компаний. Это ограничение определяет валидность вывода. Электротепловую систему нельзя оценивать числами, которые лишь притворяются данными недоступного объекта.

2. Лестница доказательств: что произошло в Саудовской Аравии

Используются пять описаний доказательств: **заявление о намерении, демонстрация приложения, объявление о работающей услуге, измерение поставщика или бенчмарк названной платформы, независимое воспроизводимое измерение на конкретной площадке**. Они не образуют простой линейной шкалы: независимый тест ускорителя может существовать в США до документированного ввода того же ускорителя в другой стране. Классификация показывает, почему заголовки нельзя свести к одному слову «достижение».

Дата	Событие	Что подтверждает источник	Чего из него нельзя заключить
12 сентября 2024 года	Партнёрство Groq и Aramco Digital по центру инференса; голосовая демонстрация; рынок pawat	Объявление о сотрудничестве, интерактивная демонстрация и оценки будущего масштаба S01	Что вся запланированная мощность установлена и приносит выручку
Декабрь 2024-го и февраль 2025-го	Установка кластера за восемь дней; объявление GroqCloud из Даммама	Сообщённое оператором эксплуатационное событие с определёнными периодом и местом S02	Электрическая мощность, расход воды, подробное число серверов или локальная выручка
Февраль 2025 года	Саудовское обязательство на 1,5 млрд долларов для расширения инфраструктуры	Финансовое обязательство по расширению S02	Фактические расходы к настоящему времени или число приобретённых стоек
Май 2025 года	Groq объявлена поставщиком инференса HUMAIN	Заявленные отношения поставки S03	Фактическая доля запросов HUMAIN и состав закупок
24 декабря 2025 года	Неисключительная лицензия NVIDIA; переход основателя и инженерной команды	Объявленное технологическое соглашение; юридическая самостоятельность Groq по её заявлению S06	Продажа всей компании или прекращение её услуг
24 августа 2026 года	Groq 3 LPX запущена в серийное производство; Nebius названа первым внедрившим провайдером	Заявленное состояние продукта, производства и внедрения S07	Что эта версия работает в Даммаме
Август 2026 года	Groq сообщает о сотрудничестве с Dell и планах внедрения LPX	План поставки и будущая цель S12	Что продукт уже установлен в Саудовской Аравии

Кажущееся географическое противоречие. В сообщениях февраля и мая 2025 года назван Даммам, современная страница groq.com перечисляет точку **SA-1 — Riyadh**. Новый источник не объясняет различие, а лишь указывает 13 центров сети и часть местоположений. Мы сохраняем обе записи, не сочиняя историю переезда и не считая их двумя объектами. Для карты кластера нужны прямые сведения оператора или документы площадки. [S02](#) [S03](#) [S08](#)

Что означает собственный чип Groq? Поставщик заявлял, что LPU и связанные системы спроектированы, изготовлены и собраны в Северной Америке. «Фабрика инференса в Саудовской Аравии» означает **место вычисления и оказания услуги**, а не владение архитектурой чипа или фабрикой кремния. Саудовская модель может работать на американском чипе в саудовском центре через глобальный коммерческий интерфейс. Для измерения локальной ценности нужно установить вклад каждого слоя, не перескакивая от одного к другому.

3. Почему скорость токенов не пропорциональна FLOPS

Первый инструмент анализа из архитектуры компьютеров — **арифметическая интенсивность** (*arithmetic intensity*) $I = F/B$, где F — число операций, а B — байты перемещаемых данных. Если вычислительная способность равна C операций в секунду, а пропускная способность памяти — M байт в секунду, упрощённый теоретический предел производительности имеет вид:

$$P_{\max} \leq \min(C, IM)$$

Этого недостаточно для расчёта реальной модели. Приближение **Roofline** исключает планирование ядер, сетевые вызовы и мелкие операции, разбиение между ускорителями, репликацию данных и синхронизацию. Но основное различие остаётся: при $IM < C$ одно увеличение арифметической мощности не поможет, пока не улучшатся перемещение памяти или повторное использование данных.

На стадии **prefill** многие позиции контекста обрабатываются параллельно, образуя плотные матричные операции. На стадии **decode** авторегрессионной модели следующий токен зависит от предыдущего. Множество пользователей позволяет объединять запросы в пакеты, но не устраняет зависимость внутри последовательности одного пользователя. Более широкая память может повысить индивидуальную скорость генерации, однако размещение весов и состояний KV остаётся задачей ёмкости и распределения.

Для проверки единиц рассмотрим нейтральную модель с **70 млрд параметров**. Минимальные сырые объёмы весов без дополнительных объектов таковы:

Представление веса	Расчёт	Только веса
16 бит	70 млрд × 2 байта	140 десятичных ГБ
8 бит	70 млрд × 1 байт	70 десятичных ГБ
4 бита	70 млрд × 0,5 байта	35 десятичных ГБ

Это **расчёт по определению, не бенчмарк**. Из него нельзя вывести необходимое число LPU: межчиповое распределение зависит от разбиения весов, KV cache, другой памяти, соединений, квантования и компилятора. Заявленные для стойки NVIDIA Groq 3 LPX 128 ГБ общей SRAM при 256 LPU не означают, что модель 70B в 16 битах должна быть отвергнута; они показывают, что **ёмкость, разбиение и плотность** не исчезают за одним показателем пропускной способности. Так же нельзя считать справедливым

сравнением задач заявленные Groq более 80 ТБ/с SRAM старого LPU и отдельные показатели HBM. Нужно зафиксировать модель, точность, размер пакета, контекст и состояние KV. [S04](#) [S05](#) [S07](#) [S08](#)

4. «Детерминизм» Groq: что гарантирует компилятор

Groq описывает архитектуру как *programmable assembly line*: компилятор делит данные и инструкции и распределяет работу между исполнительными узлами и соединениями. Многие межпроцессорные перемещения оказываются **назначены на известные моменты времени**. Это внутренняя детерминированность программных операций, а не обещание постоянной полной задержки для каждого пользователя. Управление внутри чипа сокращает один источник неопределённости во время исполнения; снаружи остаются длина запроса, очередь, конкуренция моделей, преобразование данных и речи, интернет и ограничения API. [S04](#) [S05](#) [S09](#)

При осознанном упрощении полное время задачи можно представить так:

$$T_{task} = T_{net} + T_{queue} + T_{prefill} + \sum_{k=1}^N T_{decode,k} + T_{tools} + T_{post}$$

Это **операционная сумма, показывающая состав времени**, а не предположение об одинаковом измерении компонентов: при потоковой или параллельной обработке некоторые участки перекрываются. Следует различать **TTFT, межтокенную задержку и полное время**. Первое слово может прийти быстро, а остаток фразы — медленно. Интерфейсы Prometheus Groq документируют метрики очереди, TTFT и конца запроса, вокруг которых в принципе можно строить опыт. Здесь живая система Groq не использовалась, ключей и данных саудовского клиента не было, извлечение таких метрик не заявляется. [S10](#)

В агентных системах повторение добавляет ещё один слой: короткая задача из пяти последовательных циклов генерации может сильнее зависеть от времени токена между циклами, чем длинный однократный запрос. Преимущество нужно оценивать по **времени до принятого результата задачи**, а не по скорости однопользовательского теста. Это помогает понять гибридную архитектуру NVIDIA: Rubin GPU и LPX объединены для разных фаз. Из такого решения нельзя вывести, что GPU «проиграл», а LPU «победил», или наоборот. Сам гибрид показывает, что отдельные фазы могут выигрывать от разных архитектур. [S07](#) [S12](#)

5. Выбор оптимизируемой переменной

Вместо «числа токенов» предлагается более содержательная величина выпуска Y :

$$Y = \sum_{j=1}^J \mathbf{1}\{\text{quality}_j \geq q_j\} \mathbf{1}\{\text{finish}_j \leq d_j\} v_j$$

Здесь j — задача, q_j — заранее заданный порог приемлемого качества, d_j — крайний срок, v_j — объявленный до исполнения вес ценности. Если положить $v_j = 1$, показатель Y считает принятые задачи, завершённые с требуемым качеством и в срок. В реальной оценке квантование и перенос задачи между моделями способны изменить качество. Для полевого исследования два показателя следует читать вместе:

$$\eta_E = Y/E_{site}, \quad \eta_W = Y/W_{site}.$$

Их нельзя свести к обратным PUE и WUE. Знаменатель PUE — E_{IT} , и у площадочного WUE знаменатель также E_{IT} , тогда как здесь инженерно-экономический результат — **успешная задача, завершённая в срок**. При низкой загрузке некоторые коэффициенты объекта могут быть стабильны, хотя число принятых задач в час резко меняется.

В синтетическом расчёте используется другая величина. Пусть $B_t = \sum_j x_{j,t}$ — пакетный объём, выполненный в час t . Тогда общий объём равен $W = \sum_{t=0}^{47} (U_t + B_t)$.

Интерактивные и пакетные единицы складываются с одинаковым весом на общей искусственной шкале. Поэтому $W \approx 4394,2$ — непрерывный синтетический объём нагрузки, а не число принятых задач по формуле Y . Отдельно модель завершает 48 пакетных заданий. Отношения W/E_{site} и W/W_{site} пригодны только для внутреннего сравнения политик этой модели; они не заменяют полевые показатели Y/E_{site} и Y/W_{site} .

Отдельно проверяются четыре причинных пути: (а) модель, квантование и пакетирование определяют вычисление и движение памяти; (б) вычисление превращается в тепло; (в) охлаждение и погода задают дополнительное электричество и площадочную воду; (г) источник энергии добавляет воду за пределами площадки. Нельзя сложить площадочную и электрическую воду и назвать сумму WUE без объяснения границ. Современная литература требует указывать место расхода, охлаждение, водоёмкость электричества и фактическую нагрузку; обзор Lei et al. 2025 года обнаружил очень широкий разброс водопотребления нагрузки между условиями и технологиями. [S11](#) [S13](#)

6. Физика, предотвращающая неловкие выводы: зачем теплу два названия

В дата-центре тепло проходит от транзисторного перехода к корпусу, поверхности отвода, охладителю или воздуху, теплообменнику и внешней среде. Жидкостная холодная пластина может работать в почти замкнутом контуре внутри стойки; поэтому

«жидкостное охлаждение» само по себе не означает постоянного испарительного расхода воды. В испарительной системе часть энергии уходит в пар; требуются подпитка, контроль солей, продувка и учёт потерь. S11

Два основных погодных показателя не взаимозаменяемы:

- **Температура сухого термометра T_{db}** — обычная температура воздуха, связанная с его явной теплотой.
- **Температура влажного термометра T_{wb}** — температура смоченной поверхности при испарении; она отражает способность воздуха принять дополнительный водяной пар и задаёт теоретический предел испарительного охлаждения.

При одинаковой T_{db} города могут сильно различаться по T_{wb} из-за влажности. Производительность теплообменников зависит и от сближения температуры жидкости с температурой воздуха, и от проектных ограничений. Поэтому правило «запускайте ИИ ночью» не универсально. Более прохладная ночь не всегда выгодна для испарения, не гарантирует дешёвого электричества и ничего не говорит о воде, затраченной на его производство.

Идеальная оценка фазового перехода показывает масштаб. Для испарения килограмма воды примерно при 25 °C требуется около 2442 кДж/кг скрытой теплоты, а 1 кВт·ч равен 3600 кДж. Следовательно, $3600/2442 \approx 1,474$ кг воды испаряется на 1 кВт·ч тепла при **идеальном полном испарительном отводе**. При плотности около килограмма на литр это примерно 1,47 л/кВт·ч отведённого тепла. Это не абсолютный предел воды многорежимной системы, не полный забор воды из сети и не показатель облачной услуги. S14

Связанные, но разные определения:

Символ	Определение	Единица	Концептуальная граница
E_{IT}	Энергия вычислительного оборудования, памяти и учитываемой ИТ-сети	кВт·ч	Зависит от квантования, простоя и типа ускорителя
E_{site}	Энергия объекта, включая ИТ, охлаждение и вспомогательные системы	кВт·ч	Не обязательно равна энергии первичного производства до сетевых потерь
W_{site}	Вода, потреблённая на площадке в границах опыта	л	Отличается от водозабора и воды электрогенерации
PUE	E_{site}/E_{IT}	безразмерно	Сравним только при согласованных периоде и границах
WUE	W_{site}/E_{IT}	л/кВт·ч	Здесь площадочный показатель; upstream-вода исключена
W_{elec}	Вода цепочек производства энергии, отнесённая к потреблению	л	Нужен локальный и временной коэффициент сети, которого здесь нет
Y	Принятые задачи, завершённые с нужным качеством и в срок	число или взвешенная ценность задач	Определяется заранее в полевом исследовании
W	Суммарный синтетический объём интерактивной и пакетной нагрузки	условные единицы	Искусственная общая шкала; не число задач и не метрика реального оборудования

Расширенная многоресурсная цель могла бы иметь вид $E_{site} + \lambda\{W_{site} + W_{elec}\}$. Но W_{elec} **не рассчитывается**: нет почасового предельного саудовского фактора с указанием места генерации и типа воды. Неизвестная величина не подменяется нулём, а оставляется **за границей модели**.

7. Синтетический опыт: сорок восемь часов без названия города

Чтобы отделить вычислительное доказательство от полевого, созданы два 48-часовых погодных набора с вымышленным суточным циклом, повторённым дважды. T_{db} и T_{wb} следуют синусоидальной кривой с разными параметрами для режимов hot_dry и coastal_humid. Названия лишь условно описывают два режима и не обозначают саудовские метеостанции. У нулевого часа нет календарной даты и соответствующего измерения Национального центра метеорологии. Его портал документирует исторические и текущие данные и прогнозы через API с учётными данными;

подтверждённый почасовой ряд для этой работы не получен. Среднегодовая климатическая таблица также не оправдывает заявление о реальном часе в конкретном дата-центре. [S18](#)

Затем придуманы **48 пакетных заданий** $j = 0 \dots 47$: одно поступает каждый час и должно быть завершено в пределах окна до девяти часов после доступности. На все 48 часов добавлена интерактивная нагрузка U_t , которую модель не позволяет переносить. Синтетическая энергия задана как 0,8 кВт·ч на интерактивную единицу и 1,0 кВт·ч на пакетную. Постоянные 18 кВт·ч энергии ИТ в час представляют простой и готовность, предел равен 150 кВт·ч ИТ за час. «Единица работы» не соответствует числу токенов, чипу Groq или мощности реального объекта: это абстрактная величина для анализа временного выбора.

Синтетическая дополнительная электрическая нагрузка охлаждения задаётся долей $c_{m,t}$ энергии ИТ в час t , где $m \in \{dry, evap\}$. Водный коэффициент $w_{m,t}$ измеряется в л/кВт·ч ИТ. В коде нагрузка **сухого охлаждения** растёт при высокой температуре воздуха, а **испарительное** охлаждение в некоторые часы требует меньше электричества, но больше воды. Для сухого режима задан малый ненулевой расход 0,035 л/кВт·ч ИТ на вспомогательные нужды; называть его полностью безводным нельзя.

Входной параметр	Вымышленное значение	Зачем он нужен	Чего он не означает
Горизонт	48 часов	Даёт окно планирования внутри условных суток	Не описывает сезонный год
Максимальная мощность	150 кВт·ч ИТ за час	Создаёт правдоподобное ограничение	Не соответствует коммерческой стойке
Энергия готовности	18 кВт·ч ИТ/ч	Не позволяет объявить расход нулевым после переноса работы	Не измеряет простой Groq
Интерактивная единица	0,8 кВт·ч	Нормирует срочную работу	Не равна токену
Пакетная единица	1,0 кВт·ч	Делает перенос вычислимым	Не равна метрике GPU или LPU
Гибкость пакета	До 9 часов	Ограничивает сдвиг нагрузки	Не обобщается на каждый Batch API
λ	От 0 до 0,12 кВт·ч/л	Строит математический компромисс	Не является ценой воды или официальным весом

Точные команды построения входов находятся в воспроизводимом пакете. Это важнее округлённой таблицы: читатель видит кусочные функции $\max$, моменты переключения охлаждения, может изменить параметры и повторить решение. S17

8. Математическая задача и её разрешимость

Пусть $x_{j,t} \geq 0$ — объём пакетного задания j , выполненный в час t . Переменная разрешена только между моментом доступности r_j и сроком d_j . Требуемый объём задания — b_j ; оно должно быть завершено полностью:

$$\sum_{t=r_j}^{d_j} x_{j,t} = b_j, \quad x_{j,t} \geq 0.$$

При постоянной мощности в течение часа энергетическое ограничение имеет вид:

$$18 + 0.8 U_t + \sum_{j:r_j \leq t \leq d_j} x_{j,t} \leq 150.$$

Поскольку пакетная работа делится на единицы, подходит **линейное программирование** (*linear programming, LP*). На практике задания могут быть неделимыми, перемещение KV и запуск экземпляров моделей — дискретными, а размеры пакетов — целыми. Тогда потребуется **смешанно-целочисленное линейное программирование** (*mixed-integer linear programming, MILP*) или пошаговый онлайн-алгоритм. Дробная модель не решает эти детали.

Для каждого часа до построения стоимости LP выбирается режим охлаждения с наименьшим значением $c_{m,t} + \lambda w_{m,t}$. После фиксации m_t стоимость единицы пакетной работы в этот час равна $1 + c_{m_t,t} + \lambda w_{m_t,t}$, и решается:

$$\min_x \sum_{t=0}^{47} \sum_j x_{j,t} [1 + c_{m_t,t} + \lambda w_{m_t,t}].$$

Это разделимая двухэтапная оптимизация: почасовой $\arg \min_m (c_{m,t} + \lambda w_{m,t})$ выполняется до LP, а переменными линейной программы остаются только $x_{j,t}$. Такая последовательность равносильна совместному выбору режима и распределения лишь потому, что затраты здесь линейны и разделимы по часам, а переключение бесплатно и не связано с минимальным временем работы или межчасовыми ограничениями.

Энергия интерактивной нагрузки и готовности остаётся частью **итоговой оценки** каждой политики. Она не влияет на выбор x после фиксации охлаждения, поскольку постоянна. У переключения охлаждения в модели нет стоимости и минимального времени работы — оба допущения нужно снять в промышленном исследовании.

Реальный объект может не иметь выбора между сухим и испарительным режимом в пределах часа; программа не добавит водяную башню зданию, спроектированному без неё.

Задача решается через **SciPy HiGHS / linprog** с равенством полного объёма каждого задания и почасовыми ограничениями мощности. Две базовые политики запускают работу при первой возможности (**FIFO**) с постоянно испарительным или постоянно сухим режимом. Ещё одна **FIFO**-политика каждый час выбирает охлаждение с наименьшим электричеством; оптимизированная политика сначала выбирает почасовой режим указанным независимым минимумом, а затем переносит задания внутри допустимых окон. Для анализа чувствительности решение повторяется при **восьми весах** воды от нуля до 0,12; пакет включает все точки. В энергетических функциях нет модели GPU или LPU, поэтому результат не сравнивает чипы. S17

9. Результаты исполнения и внутренняя проверка

Программа фактически исполнена локально с SciPy; документация, входные данные и результаты входят в пакет воспроизведения. Таблица выбирает **семь политик** для каждого погодного режима, всегда на одном периоде и одной нагрузке. **FIFO** — минимум энергии запускает пакеты при первой возможности и выбирает менее энергоёмкое охлаждение в каждый час; **LP**, λ означает линейную оптимизацию с заданным весом воды. Значения зависят от вымышленных функций и не имеют статистического ранга вне модели.

Синтетическая погода	Политика	Общая энергия, кВт·ч	Вода на площадке, л	Часы сухого охлаждения
Жаркая сухая	FIFO — испарительное	5223,07	6818,11	0
Жаркая сухая	FIFO — сухое	5490,52	165,20	48
Жаркая сухая	FIFO — минимум энергии	5192,71	4829,34	18
Жаркая сухая	LP, $\lambda=0$	5179,88	4166,87	18
Жаркая сухая	LP, $\lambda=0,02$	5188,14	3559,52	22
Жаркая сухая	LP, $\lambda=0,05$	5243,32	2222,32	30
Жаркая сухая	LP, $\lambda=0,12$	5407,56	165,20	48
Влажная прибрежная	FIFO — испарительное	5277,24	7076,90	0
Влажная прибрежная	FIFO — сухое	5504,52	165,20	48
Влажная прибрежная	FIFO — минимум энергии	5261,37	5451,87	14
Влажная прибрежная	LP, $\lambda=0$	5245,08	4893,67	14
Влажная прибрежная	LP, $\lambda=0,02$	5263,52	3640,80	22
Влажная прибрежная	LP, $\lambda=0,05$	5313,90	2248,73	30
Влажная прибрежная	LP, $\lambda=0,12$	5438,69	165,20	48

Внутренняя проверка. Каждая политика обработала один и тот же синтетический объём $W \approx 4394,2$ условной единицы, складывая интерактивную и пакетную нагрузку на общей искусственной шкале. Все 48 пакетных заданий завершены; это отдельный счётчик и не основание называть W числом задач. Нарушений момента доступности, сроков и предела ИТ-мощности — ноль. Код проверяет суммы стоимости, равенство частей задания требуемому объёму, принадлежность каждой части допустимому окну и отвергает неуспешный статус решателя. Эти проверки подтверждают **целостность исполнения записанной модели**, а не реалистичность тепловых коэффициентов, поведение клиента или оборудование Groq.

Переход от LP, $\lambda=0$ к LP, $\lambda=0,02$ в жарком сухом режиме повышает энергию на **8,26 кВт·ч** и снижает воду на **607,35 л** за 48 часов. Во влажном прибрежном режиме энергия растёт на **18,43 кВт·ч**, площадочная вода сокращается на **1252,87 л**. Это точные результаты вымышленной программы; проценты нельзя представлять экономической или экологической отдачей саудовского плана. Сравнение показывает методологический урок: **разные коэффициенты охлаждения меняют количество сэкономленных литров, покупаемых дополнительным киловатт-часом.**

Рассчитанная граница компромисса для двух вымышленных погодных наборов

Рисунок 1. Каждая точка получена фактическим решателем, значение λ раскрыто. Ни одна точка не является полевым измерением или результатом реальной электросети.

10. Почему вычислительная игра работает и всё же недостаточна для машины

В теории перенос пакета выглядит бесплатной заменой одного часа другим. На практике польза может исчезнуть в четырёх местах.

Во-первых, статическая энергия и выключение. Если стойки сохраняют большую долю мощности в простое, перенос почти не снизит электричество. При глубоком выключении пробуждение не мгновенно: модели могут загружаться и заполнять память несколько минут. Модель учитывает постоянную готовность и не сводит энергию к сумме заданий, но не исследует выключение, запуск и резерв под всплеск интерактивного спроса.

Во-вторых, резидентность моделей в SRAM/HBM. Перенесённый пакет может столкнуться с пределом памяти. Модель с большим числом весов и KV распределяется между чипами; выгрузка одного состояния и восстановление другого способны съесть выигрыш охлаждения. Вместе с энергией следует измерять *model residency*, *KV cache occupancy*, *swap traffic*, *compilation cache* и *interconnect load*.

В-третьих, уровень обслуживания и смесь моделей. Пакетная очередь может не быть чувствительной ко времени сама по себе, но интерактивный пользователь способен конкурировать с ней за то же оборудование. Хвост TTFT и времени завершения должен укладываться в срок каждого класса, а не только в среднее. Разброс длин контекста и ответа меняет спрос prefill и decode, поэтому реальный опыт должен воспроизводить распределение длин и волн поступления.

В-четвёртых, многоконтурная термодинамика. Эффективность зависит от площади поверхностей, теплообменника, энергии перекачки, температуры входа серверов, дождя, пыли и обработки подпиточной воды, а не только от наружного термометра.

Ненаучно предполагать почасовое переключение реального объекта между сухим и испарительным режимом без спецификации. Полевой опыт должен включить кривые частичной нагрузки (*part-load performance*), температурный напор охлаждения (*cooling approach temperature*), кривые вентиляторов и насосов и стоимость смены режима.

Здесь проходит эпистемическая граница между **доказательством возможности** (*existence proof*) и **оценкой эффекта** (*effect estimate*). Первое показывает, что при поддерживающих коэффициентах гибкость математически превращается в многоресурсную оптимизацию. Второе требует доказать соответствие коэффициентов реальным машинам, погоде и нагрузкам и честно сравнить с текущей системой на одинаково распределённых заданиях и измерениях.

11. Предшествующие исследования: что было сделано раньше

Научная задача возникла не из подсказки этого проекта. Обзор Беркли 2025 года изучает связь эффективности серверов, электричества, воды электросети, охлаждения и климата. Министерство энергетики США представляет PUE и WUE как полезные отношения с оговорками о границах; различие площадочных и внешних ресурсов известно задолго до этого исследования. [S11](#) [S13](#)

Препринт **Balancing Bits and Drops: Stress-Adjusted Water Management for Data Centers** 2026 года рассматривает пространственный и временной дефицит воды и планирование нагрузки с учётом воды, сухого охлаждения и внешнего следа. Поэтому утверждение, будто здесь создан «первый водный планировщик», было бы ложным. Предлагаемый вклад уже: **изучение многофазного инференса на саудовском ресурсе с разделением классов заданий, телеметрии ускорителя, охлаждения и времени.** [S15](#)

Препринт **PowerSlider**, опубликованный в августе 2026 года, изучает мгновенные ограничения мощности и учитывает асимметрию prefill/decode и сроки качества обслуживания в краткосрочном повторяющемся решении. Авторы сообщили сохранение goodput на уровне 78,3% при снижении мощности на 30% против 47,6% у сильнейшей базовой линии в своей среде. Эти числа нельзя переносить на LPU, Groq или Саудовскую Аравию; электрическая цель работы не совпадает с целью площадочной воды. Однако она показывает, что **фазово-осознанное планирование мощности** имеет исполняемую систему, нагрузки и опубликованные опровержимые эксперименты. [S16](#)

Заявления поставщиков также дают проверяемые метрики. NVIDIA описала результат Gemma 4 31B на Groq 3 LPX и высокую производительность в агентной работе. Корректная ссылка привязывает показатель к платформе, модели, контексту и

условиям бенчмарка, не превращая его в среднее для саудовских услуг. Справедливое сравнение должно различать токены одного пользователя, совокупные токены и полное время задачи. S07

12. Саудовский полевой опыт, способный опровергнуть гипотезу

Этот раздел — **предложение, которое не исполнено**. Он отличается от описанного выше синтетического расчёта. Потребуется разрешение оператора вычислительного центра, сертифицированные электрические и водные измерения и возможность воспроизводить нагрузки без риска для пользовательских данных и производства. Публичное раскрытие частной архитектуры необязательно: достаточно контролируемых статистических агрегатов, позволяющих понять метод и проверить уравнения.

Выборка и страты. В течение нескольких недель собирается эксплуатационный след, охватывающий разные температуру, влажность и длины запросов. До начала анализа работа делится на три класса: живой разговор с жёсткой задержкой; многошаговые агентные задачи с разумным сроком завершения; пакетная работа с реальными крайними сроками. Во всех экспериментальных политиках сохраняются распределения входов и выходов, временные метки и версии моделей. Частные тексты исключаются либо преобразуются в анонимные метрики длины после ясного определения сравнимых признаков.

Группы сравнения. (а) текущий производственный планировщик; (б) планирование с учётом сроков и загрузки GPU/LPU; (в) планирование с учётом времени и электричества; (г) планирование с учётом времени, электричества и площадочной воды; (д) вычислительная группа с добавлением фактора воды электрогенерации, если появится достоверный источник. Все группы подчиняются одинаковым пределам качества, обслуживания и доступной мощности. Если опыт начинает вредить реальному запросу, включается защитный механизм. Последняя группа не считается лучшей заранее: если полезная работа на единицу ресурса не улучшится в пределах доверительных интервалов, гипотеза отвергается.

Аппаратный слой. GPU общего назначения сравнивается с LPU или гибридной платформой только при **одинаковых моделях и сопоставимых точности и качестве**. Документируются численная точность (*precision*), квантование, длина контекста, ёмкость KV, размещение моделей, число одновременных пользователей, сеть и последовательность инструментов. Нельзя сравнивать веб-демонстрацию одного

ускорителя с производственной средой другого в другой сети и приписывать разницу одной физике. Если модель не запускается на чипе с тем же качеством, это важный результат совместимости, а не строка для удаления.

Экологический слой. Одновременно собираются T_{db} , T_{wb} , температуры подачи и возврата жидкости, мощность вентиляторов и насосов, счётчик подпитки, **водозабор и потребление отдельно**, источник и тип воды при доступности. Электричество ИТ, PDU, UPS и охлаждения считывается при явно заявленных границах. Публикуются точность приборов, интервалы выборки и несовпадение часов энергии и запросов, чтобы задержанные измерения не создавали ложную временную связь.

Критерии успеха. Главные показатели — число правильных задач, завершённых в срок; энергия объекта и площадочная вода на тысячу действительных задач. PUE и WUE приводятся параллельно, как и TTFT и время завершения на p50/p95/p99. Обязательно сообщаются потери качества и повторы: ускоренная выдача неверного ответа не означает роста производства. В расчёт входят время и эксплуатационная стоимость подготовки и компиляции нескольких моделей; рабочий день нельзя свести к одному быстрому токenu.

Границы причинного вывода. Сравнения прохладной недели с жаркой недостаточно: одновременно меняются смесь запросов, версии ПО, погода, доступная мощность и цены энергии. Лучше случайно чередовать разрешённую тестовую нагрузку в близких по условиям окнах, блокировать перенос эффектов предыдущей группы — например, прогрев стойки и загрузку модели — или использовать перекрёстный дизайн (*randomized crossover*) с периодами очистки. Исследователь должен заранее опубликовать распределение исходов, критерии решения, отрицательные результаты и различия условий, прежде чем делать вывод для Королевства или проектов стоимостью в миллиарды.

13. Следующая экономика: от стоимости процессора к ценности преобразования электричества во время

Город с вычислительной инфраструктурой может продавать паспортную мощность, сырые выходные токены или задачи с обязательным качеством и сроком. Эти продукты передают разную долю промышленной компетенции. В первом случае продаётся оборудованная ёмкость; во втором — более мелкая счётная единица. В третьем нужны инженерия длины контекста, сопоставление модели с ускорителем, сеть, планировщик, счётчики энергии и воды и опыт, доказывающий полезность результата. Лучшее распределение нагрузки способно увеличить выпуск без дополнительных ускорителей, а стоимость реализации и сложность способны поглотить весь выигрыш. Это не прогноз выручки, а **проверяемая микроэкономическая гипотеза**.

Рядом с исходной услугой Groq для Саудовской Аравии возникает операционная ниша: **преобразователи нагрузок инференса между разнородными архитектурами**, планировщики с локальными измерениями и библиотеки бенчмарков, связывающие модель и контекст с качеством на арабском и других языках. Такие платформы могли бы измерять **принятые задачи на джоуль** (*accepted tasks per joule*) и **принятые задачи на литр воды площадки** (*accepted tasks per site liter*), отделяя работу, которой нужны доли секунды, от работы с несколькими часами гибкости. Но продажи и защищённое преимущество не следуют из общего роста инвестиций. Ценность решений ПО нужно доказать против собственных инструментов NVIDIA и Groq, университетских работ и систем поставщиков охлаждения, иначе новое название лишь прикроет существующий инструмент.

Период 2027–2035 годов здесь представляет не численный прогноз, а дерево проверок. Если появятся пригодные стандартные данные охлаждения, а разнородные ускорители будут работать с ограниченной ценой переноса, физическое планирование может войти в **контур управления** (*control plane*) услуги инференса. Если большая часть нагрузки останется мгновенной или перемещение состояния будет дороже временного выигрыша, разумнее снижать энергию токена через компилятор, квантование и электрическую эффективность. Если площадочная экономия воды конфликтует с водой внешней электрогенерации, победителя нельзя объявить до фиксации границ учёта. Эти условные выводы должны меняться под воздействием опыта, а не объявляться фактами будущего.

14. Вычислительная воспроизводимость и независимость

Сопровождающий пакет воспроизведения содержит код, входы и проверяемые результаты. После распаковки расчёт и графики воспроизводятся командами:

```
python3 analysis/reproduce.py
python3 analysis/plot_figures.py
```

- Скачать код, данные и результаты воспроизведения — S17.

Расчёт использует **NumPy**, **SciPy** и **Matplotlib**, без ключей API и внутренних вызовов частного продукта. Входы — `data/weather_synthetic.csv` и `data/workload_synthetic.csv`; подробные результаты — `data/policy_hourly.csv`, `data/policy_summary.csv` и `data/policy_sensitivity.csv`; параметры запуска — `data/run_manifest.json`. `evidence_workbook.xlsx` содержит проверяемую таблицу результатов, формулы и ссылки на входы. Иллюстрации сохраняются в PNG и SVG, а реестр явно помечает их синтетическую природу.

Проверочный запуск воспроизвёл все показанные агрегаты и разности с заявленной точностью округления. Единственное почасовое распределение, однако, не гарантировано: в файле зависимостей указаны лишь минимально допустимые версии, а у линейной задачи есть равнозначные оптимумы. В проверке 48 из 1056 почасовых строк получили иное допустимое распределение пакетной нагрузки, сохранив те же целевые суммы, объём W , число режимов охлаждения, предел мощности и соблюдение сроков. Чтобы всегда получать одно и то же почасовое решение, нужно закрепить версии Python, NumPy, SciPy и HiGHS и добавить вторичную цель или правило разрешения ничьих.

В статье нет **вероятностной выборки** климата или компаний и нет оправданного доверительного интервала для промышленного эффекта. HiGHS решает записанную линейную задачу; корректность решения в рамках допущений не делает его национальной или поведенческой оценкой. Реальные очереди саудовских центров не тестировались, расход на токен чипа Groq не подтверждался. Читатель может повторить код, заменить кривые разрешёнными измеренными зависимостями и проверить различия. **Вычислительная воспроизводимость** (*computational reproducibility*) и **независимое эмпирическое повторение** (*empirical replication*) принципиально различаются: пакет обеспечивает воспроизводимость агрегатов с показанной точностью, но не единственность почасового оптимума; эмпирическое повторение остаётся будущим научным проектом.

15. Чего мы не знаем и чего график не должен изображать

Публичные источники не раскрывают потребление электричества и воды центром Groq в Даммаме, конструкцию отвода тепла, тип и число чипов на каждой площадке, количество токенов, произведённых только в Саудовской Аравии, или водную мощность, связанную с инвестиционным обязательством. Не найдено доказательств установки NVIDIA Groq 3 LPX на конкретном саудовском объекте, хотя Groq объявила планы раннего внедрения. Не показано независимое преимущество LPU над GPU на одинаковой задаче внутри Королевства. Статус «поставщика инференса HUMAIN» — коммерческое заявление, а не реестр счетов; из него нельзя вывести объём использования или выручку.

Почасовые данные NCM не получены. Слово «прибрежный» в опыте обозначает вымышленный погодный профиль. WUE реального центра, водный след опреснения и вода электросети не измерены. У объекта может не быть режима охлаждения, предполагаемого моделью, или переключение может запрещаться эксплуатационными ограничениями. Это не недостатки стиля, которые исправляются дополнительными рекомендациями, а точные границы доказательства.

Опровержимый вывод. В Саудовской Аравии есть реальный заявленный эксплуатационный сигнал Groq и возможность проверить ПО, которое увеличивает **правильную работу, выполненную в срок, на единицу электричества и воды**. Научный и экономический успех нельзя начинать с предпосылки, что LPU лучше во всём, или с обещания, что сухое охлаждение само по себе означает устойчивость. Начать следует с двух эксплуатационных вопросов: **что действительно ограничивает скорость задачи — память, вычисление, сеть или очередь? Сколько энергии и воды потребовала эта задача в этом месте и в этот час по проверяемым счётчикам?**

Ссылки и происхождение доказательств

Ниже перечислены первичные материалы организаций и исследовательские работы. Маркетинговое заявление не считается независимой проверкой.

- [S01 — Groq / Aramco Digital, объявление о партнёрстве, 12 сентября 2024 года](#). Подтверждает меморандум, голосовую демонстрацию и будущие цели; не подтверждает фактический объём услуги.
- [S02 — Groq, работа в Даммаме и саудовское обязательство расширения, 10 февраля 2025 года](#). Обязательство не равно подтверждённым расходам; электричество, вода и токены центра не раскрыты.
- [S03 — Groq, статус поставщика инференса HUMAIN, 27 мая 2025 года](#). Заявленная связь поставки не подтверждает объём или выручку.
- [S04 — Groq, объяснение LPU, 7 марта 2025 года](#). Описывает статический поток данных и SRAM; показатели скорости и энергии остаются зависящими от поколения заявлениями поставщика.
- [S05 — официальный технический лист GroqChip Processor, поколение 2024 года](#). Это документ поставщика о соответствующем поколении; его нельзя подменять данными Groq 3 LPX.
- [S06 — Groq и NVIDIA, неисключительная лицензия на технологию инференса, 24 декабря 2025 года](#). Полные условия лицензии публично не раскрыты.
- [S07 — NVIDIA, Groq 3 LPX в серийном производстве, 24 августа 2026 года](#). Бенчмарк относится к заявленной платформе и тесту, а не к подтверждённому саудовскому развёртыванию.
- [S08 — платформа Groq и сеть из 13 центров](#). Метка Riyadh не объясняет более раннее упоминание Даммама и не содержит инвентаризации площадки.
- [S09 — документация Groq по уровням обслуживания и пакетным API](#). Уровни услуги не доказывают существования внутреннего водо-осознанного планировщика.

- S10 — метрики задержки Prometheus в Groq. Документация не заменяет фактическую телеметрию клиента.
- S11 — Министерство энергетики США, эффективность использования воды для охлаждения дата-центров, 9 января 2019 года. Примеры США не являются данными Саудовской Аравии.
- S12 — Groq о будущем внедрении NVIDIA Groq 3 LPX и Vera Rubin. План внедрения не подтверждает установку в Саудовской Аравии.
- S13 — Lei et al., *The water use of data center workloads*, 2025, DOI [10.1016/j.resconrec.2025.108310](https://doi.org/10.1016/j.resconrec.2025.108310). Разброс относится к условиям исследования, а не к национальному саудовскому коэффициенту.
- S14 — Национальное бюро стандартов США / NIST, измерения скрытой теплоты воды около 25 °C. Идеальная оценка фазового перехода не является WUE объекта или полным расходом подпиточной воды.
- S15 — Talukder et al., *Balancing Bits and Drops*, arXiv:2607.22617, 2026 год. Методика препринта здесь не воспроизводилась и не проверялась на саудовских объектах.
- S16 — Li et al., *PowerSlider: Exploiting Phase Asymmetry*, arXiv:2608.21719, 2026 год. Результаты относятся к установке SGLang авторов, а не к Groq, LPU или саудовской услуге.
- S17 — пакет воспроизведения: 48-часовая линейная программа, синтетические данные и результаты. Все входы вымышлены; нет наблюдений погоды, измеренных чипов, калибровки или внешней валидности.
- S18 — Национальный центр метеорологии Саудовской Аравии, начало работы с погодным API. Доступность почасовых данных документирована, но аутентифицированный ряд в исследовании не получен.