



探索性研究论文

编辑日期：2026 年 10 月 23 日

从词元工厂到热调度器

语言处理单元（Language Processing Unit, LPU）架构、沙特推理基础设施与可复现的水—电—工作权衡实验

English title: From Token Factories to Thermal Schedulers: LPU Architecture, Saudi Inference Infrastructure, and a Reproducible Water – Power – Work Trade-off Experiment.

研究类型：概念与计算工作论文（conceptual and computational working paper）。论文包含一项已实际运行、但输入完全合成的线性规划实验，并审阅原始来源与公开运营案例。它不是对 Groq 设施的现场研究，不是对其芯片性能的独立基准测试，也不是沙特国家用水量估算。

摘要。本文考察一个运营问题：沙特的语言模型推理管理，能否从“需要多少加速器”的单一决策，转为对有价值计算工作、时间、电力和冷却用水的联合优化。现实起点包括 Groq 2024 年宣布与 Aramco Digital 合作、2025 年宣布从达曼提供 GroqCloud 服务，以及采用片上静态随机存取存储器（static random-access memory, SRAM）和静态数据流调度的 LPU 架构；此后，公司于 2025 年 12 月向 NVIDIA 非独家许可技术，Groq 3 LPX 平台又在 2026 年 8 月进入全面生产。本文把运营事实、扩张计划和供应商基准主张分开处理。我们建立一个运营模型，区分模型输入处理（prefill）与串行生成（decode）、即时请求与有截止期限的批处理工作，也区分场内用水与电网供电相关的场外用水。实验使用两个合成天气情景和 48 小时时窗；在成本线性可分、冷却方式可逐时自由切换的假设下，程序先逐时选择目标成本最低的冷却方式，再由线性规划重新分配弹性工作。所有解的合成工作量 W 均约为 4394.2 个任意单位，而不是 4394.2 项任务；48 个批处理任务均在截止期限内完成。在合成的炎热干燥情景中，提高目标函数的用水权重，相较最低电耗解减少约 607 L 用水，代价是增加约 8.26 kWh 电。这个结果证明多资源优化器在程序上可以构建，不是设施实测效应，也不能证明 LPU 优于其他架构。最后，本文提出一个可被证伪的现场实验设计，并不预设热感知调度一定有效。

Abstract. This working paper separates three evidence layers: Groq’s publicly reported Saudi inference operations, the technical claims of its deterministic SRAM-centric Language Processing Unit, and a local optimization exercise with fully synthetic inputs. Under linear separable costs and costless, unconstrained hourly mode switching, the program first selects dry or evaporative cooling by hourly argmin, then uses a 48-hour linear program to allocate 48 flexible jobs with release/deadline constraints alongside fixed interactive work. Every policy produces the same arbitrary synthetic work volume W , approximately 4394.2 units; this is not the task-count metric Y , while all 48 batch jobs meet their deadlines. The experiment demonstrates a conditional electricity – on-site-water frontier, not observed savings in Saudi Arabia, Groq data centers, or LPU hardware. Its main research proposition is to integrate phase-aware serving metrics, physical cooling telemetry, and completed-work quality into an auditable Saudi inference measurement protocol. Vendor benchmarks and national announcements are qualified by their operational and epistemic boundaries. Site consumption, water withdrawal, electricity-related upstream water, real chip energy, network latency, and life-cycle impacts require independent facility data before any generalization.

关键词： Groq；语言处理单元（LPU）；静态数据流；SRAM；HBM；预填充 / 解码；KV cache；湿球温度；PUE；WUE；水资源消耗；推理；线性规划；批处理调度；沙特阿拉伯。

1. 可研究的问题与主张边界

核心问题是：如果沙特的一家推理运营商同时面对必须即时完成的工作与可延迟工作，并拥有关于负载用电、冷却方式和天气影响的可靠数据，它能否在不违反时间和质量要求的前提下，降低每项正确完成任务的用水或用电？在真实设备和真实负载的响应尚未测得之前，科学上不能直接回答“能”。因此，本文提供三个主张更窄、也更容易审计的产出：

1. 从 Groq 自身公告重建其沙特案例，标注证据类型，并区分公开观察到的事件、协议、路线图和供应商测试。
2. 建立一个因果与技术模型，连接输入长度、生成方式、存储器、调度器、服务器用电、排热、天气与用水。
3. 在合成数据上运行计算实验，展示可能的权衡空间与外推失败点，再提出用真实设施测量替换构造函数的方案。

本文不假定每词元成本更低就等于每项任务资源更省，也不假定省电必然与省水同时发生；更不假定使用专用芯片，就会在所有神经网络模型上占优。48 小时实验不属于 Groq 或 HUMAIN，它为本研究单独构造，不复刻两家公司任何已知架构。这个限制是结论成立的条件：在无法接触真实系统时，不能用伪装成设施实测的数据来验证电力与热管理优化。

2. 证据阶梯：沙特案例中实际发生了什么

本文使用五类证据描述：意向公告、应用演示、服务运营公告、供应商测量或指定平台基准、特定地点内可复现的独立测量。它们并非一条单向阶梯；公司可能先在美国取得某款加速器的独立测试，再在其他地点记录部署。这个分类说明了为什么不能把所有新闻标题都归为同一种“成就”。

日期	事件	来源支持的描述	不能据此推断的内容
2024 年 9 月 12 日	Groq 与 Aramco Digital 宣布推理中心合作、展示语音对话，并提出 nawat 市场构想	合作公告与交互演示；未来规模估计 S01	所有规划容量已安装或已经产生收入
2024 年 12 月及 2025 年 2 月	八天安装集群；宣布从达曼提供 GroqCloud S02	运营商公布的、具有时间与地点信息的服务运营事件	电力容量、用水量、详细服务器数、当地收入
2025 年 2 月	沙特承诺投入 15 亿美元扩建基础设施 S02	用于扩建基础设施的资金承诺公告	截至目前的实际支出或已采购机架数
2025 年 5 月	Groq 宣布成为 HUMAIN 推理服务提供商 S03	公布的供应关系	Groq 实际承担的 HUMAIN 请求比例及每笔采购类型
2025 年 12 月 24 日	Groq 向 NVIDIA 非独家许可技术；创始人及工程团队转入 NVIDIA S06	公布的技术许可协议；按公告，Groq 法律实体仍独立	NVIDIA 已收购整家公司，或 Groq 服务已经结束
2026 年 8 月 24 日	NVIDIA Groq 3 LPX 全面投产；Nebius 成为首家采用者 S07	公布的产品、生产与采用状态	这个版本已在达曼中心运行
2026 年 8 月	Groq 宣布与 Dell 合作部署 LPX，并计划采用该平台 S12	供应安排与未来部署目标	在资料截止时已经安装于沙特

地点信息存在表面矛盾。2025 年 2 月和 5 月的公告写的是“达曼”，Groq 当前平台页则列出 SA-1 — Riyadh（利雅得）。新页面只列举全球 13 个中心和部分地点，没有解释运营层面的差异。本文同时记录两组来源，不拼接迁移故事，也不擅自把它们画成两个中心。要厘清集群地理结构，需要运营方的直接说明或设施文件。[S02](#) [S03](#) [S08](#)

Groq 自行设计芯片意味着什么？供应商称，LPU 及相关产品在北美设计、制造和组装。因此，“沙特的推理工厂”指的是计算与服务的发生地点，不等于沙特拥有芯片设计技术或硅制造工厂。一套沙特开发的阿拉伯语模型，可以运行在美国芯片上、由沙特境内的数据中心执行，并通过全球商业接口提供服务。评估本地价值，需要逐层确定究竟在哪一层完成了什么，而不能从一层跳到另一层。

3. 为什么词元速度不与 FLOPS 成正比

计算机体系结构的第一个分析工具是算术强度（arithmetic intensity），记为 $I = F/B$ ，其中 F 是运算次数， B 是取回数据所需的字节数。若计算能力为每秒 C 次运算，存储器带宽为每秒 M 字节，则简化后的理论性能上限为：

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

这个公式不足以计算真实模型性能。它是屋顶线模型（Roofline model）的近似，省略了内核调度、网络调用、小型算子、处理器间离散分布、数据复制和同步时间。但它揭示的区别很关键：当 $IM < C$ 时，仅提高算术乘法能力不会起作用，除非数据移动或复用同时改善。

在预填充阶段（prefill），多个上下文位置可以并行处理，形成密集矩阵计算；在自回归模型的解码阶段（decode），下一个词元依赖前一个词元。多用户并发可以让请求聚合成批，却不能消除单个用户序列内部的依赖。更高的存储器带宽可能提高单用户词元生成速度，但怎样放置全部模型参数与键值缓存（key-value cache, KV cache），仍然是容量分配和管理问题。

为了核对单位，设一个中性的 700 亿参数模型。只计算权重、尚未加入其他对象时，其最低原始体积为：

权重表示	计算	仅权重体积
16 位	700 亿 $\times$ 2 字节	140 GB（十进制）
8 位	700 亿 $\times$ 1 字节	70 GB（十进制）
4 位	700 亿 $\times$ 0.5 字节	35 GB（十进制）

这是定义性计算，不是基准测试。它没有告诉我们服务一个模型需要多少 LPU，因为芯片间部署取决于权重分片、KV、其他存储管理、互连速度、量化与编译器。NVIDIA Groq 3 LPX 机架公布的配置是 256 个 LPU、合计 128 GB SRAM；这不表示 70B、16 位模型必定无法运行，而是说明容量、分布与密度不会因为一个带宽数字而消失。同样，不能把 Groq 对较早一代 LPU 公布的超过 80 TB/s 带宽，与某些 HBM 带宽直接当作两台设备完成同一任务的公平比较。比较前必须固定模型、精度、批量、上下文长度与 KV 状态。[S04](#) [S05](#) [S07](#) [S08](#)

4. Groq 所说的“确定性”：编译器保证什么，不保证什么

Groq 把其架构描述为“可编程装配线”（programmable assembly line）：编译器拆分数据和指令，把工作分配到执行单元与传输链路，处理器之间的许多消息移动因而可以按照已知时刻规划。这种确定性发生在程序操作的内部层级，不是对所有用户端到端响应时间保持不变的承诺。片上控制减少了一类运行时临场决策，系统外仍有多项变量：问题长度、服务队列状态、多模型争用、数据转换、语音处理、互联网与 API 请求限制。[S04](#) [S05](#) [S09](#)

在有意简化的前提下，一项任务的总时间可以写为：

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

这是一个用来说明时间构成的运维加法式，不假设所有分量都以相同方式测量；流式传输或并行执行时，部分时段可能重叠。必须区分首词元时间（time to first token, TTFT）、词元间延迟（inter-token latency）与端到端延迟（end-to-end latency）。最后一项包括完整输出；首个词可能很快到达，剩余句子仍可能拖延。Groq 的 Prometheus 接口文档列出了队列、TTFT 和请求结束指标，原则上可围绕它们设计实验。本文没有调用 Groq 在线接口，没有取得密钥或沙特客户数据，也没有声称抽取过这些指标。[S10](#)

智能体系统（agents）还有重复交互的额外负担：一次包含五轮连续生成的短任务，可能比单轮长请求更受轮次之间词元时间影响。优势应按经确认任务的完成时间评估，而不是只看单用户生成速率。这也有助于理解 NVIDIA 的架构决定：按其描述，Rubin GPU 与 LPX 共同处理不同推理阶段。不能把公告简化为“GPU 失败、LPU 获胜”或反过来；混合架构本身说明，对某一阶段的优化可能需要多种架构。[S07](#)
[S12](#)

5. 确定真正应该优化的变量

本文提出一个比“词元数”更严格的输出单位 Y ：

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

其中， j 表示任务， q_j 是预先确定的质量阈值， d_j 是截止时间， v_j 是运行前已声明的价值权重。当 $v_j = 1$ 时， Y 计算的是按时完成且质量合格的任务数。本文的计算程序没有直接计算这个任务计数，而是报告另一个量： $W = \sum_t (U_t + \text{batch}_t)$ 。其中， U_t 是逐时交互工作量， batch_t 是逐时批处理工作量；两者使用共同的任意合成工作单位。程序得到 $W = 4394.170722$ ，文中取约 4394.2，并另行记录 48 个批处理任务是否按时完成。因此， W 不是任务数，也不等于 $v_j = 1$ 时的 Y 。这个合成尺度高度简化；现场研究必须单独定义质量阈值与任务价值，因为权重量化或在模型间转移任务都可能改变回答质量。真实运营中，建议同时读取两个概念性指标：

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

它们不能只被处理成 PUE 和 WUE 的倒数。PUE 的分母是 E_{IT} ，按场内口径定义的 WUE 分母同样是 E_{IT} ；这里的经济与工程目标，分母则是在时限内成功完成的任务。服务器利用率下降时，某些设施比率可能保持稳定，但每小时合格任务数会显著变化。

需要分别检验四条因果路径：（a）模型选择、量化与批处理方式决定计算和数据移动量；（b）计算转化为热；（c）冷却类型与天气决定额外用电和场内用水；（d）电力来源引入场外用水。把场内用水与发电用水直接相加后仍称为标准口径的 WUE，会掩盖核算边界。现有水资源研究建议明确用水地点、冷却方法、电力耗水系数与实际负载。Lei 等人的 2025 年综述发现，不同条件和技术下，工作负载层面的用水估算差异极大。[S11](#) [S13](#)

6. 避免概念混淆的物理学：为什么温度要有两个名字

数据中心的热量经过多层传递：从晶体管结区到封装，再到散热表面、冷却液或空气、换热器，最后进入外部介质。液冷板可以把液体回路基本封闭在机架内，因此“液冷”本身不表示持续发生蒸发耗水。若采用蒸发冷却，部分能量通过水蒸气带走；补充水会被消耗，同时还要处理盐分、排污、渗漏和其他损失。[S11](#)

两个基本气象变量并非同义：

- 干球温度 (T_{db})：空气的显热状态，通常就是天气报告中的温度。
- 湿球温度 (T_{wb})：湿润表面蒸发时温度计读数，与空气接收更多水汽的能力相关；接近湿球温度时，理论蒸发冷却潜力受到限制。

两个城市的 T_{db} 可以相同， T_{wb} 却因湿度而差异很大。换热器性能也会随液体温度接近空气温度的程度和设计限制而改变。因此，“在夜间运行人工智能任务”不是普遍成立的规则。较低的夜间气温未必更利于蒸发，不保证电价更低，也没有告诉我们发这些电消耗了多少水。

理想相变计算可以说明量级。在约 25° C 时，蒸发 1 kg 水需要约 2442 kJ/kg 潜热；1 kWh 等于 3600 kJ。因此， $3600/2442 \approx 1.474$ ：在理想、完全依靠蒸发排热的条件下，每移除 1 kWh 热量约蒸发 1.474 kg 水。若密度近似为 1 kg/L，即约 1.47 L/kWh。这不是多机制冷却系统的绝对用水下限，不等于电网总取水量，也不等于一项云服务的用水量。[S14](#)

相互关联但不同的定义：

符号	定义	单位	概念边界
E_{IT}	IT 边界内计算、存储与合格网络设备的能耗	kWh	受量化、空闲状态与处理器类型影响
E_{site}	包括 IT、冷却与辅助设施在内的场内总能耗	kWh	不一定等于考虑输配损耗前的发电量
W_{site}	按实验定义计入的场内耗水	L	与取水量及发电用水分开
PUE	E_{site}/E_{IT}	无量纲	必须在一致时间范围与计量边界内比较
WUE	W_{site}/E_{IT}	L/kWh	此处为场内指标，不含上游用水
W_{elec}	归因到用电量的发电供应链耗水	L	需要分地点、分时段电网因子，本文没有该数据
Y	在质量与时限内成功完成的有用任务	价值或工作单位	现场研究必须预先定义

采用更宽核算边界时，多资源目标函数可以写成 $E_{site} + \lambda\{W_{site} + W_{elec}\}$ 。但本文没有计算 W_{elec} ：目前缺少按小时、发电地点与水类型确定的沙特边际发电耗水因子。本文没有把它当作零，而是明确放在模型边界之外。

7. 合成实验设计：没有城市名称的 48 小时

为了区分计算证明和现场证明，我们生成两组 48 小时天气数据，每组包含重复两次的虚拟日周期。 T_{db} 与 T_{wb} 都按单一相位的正弦曲线变化，参数分别对应 `hot_dry` 和 `coastal_humid`。这些名称只是两个情景的直观描述，不代表沙特气象站。小时是计算时段；第 0 小时没有对应的日历日期，也没有对应的国家气象中心（NCM）读数。NCM 门户说明其 API 可提供历史、当前和预报天气，但需要账户授权；本研究没有取得经认证的逐时数据。年度气候平均值也不足以支持对某座数据中心某个小时的主张。[S18](#)

实验构造 48 个批处理任务（ $j = 0 \dots 47$ ），每小时释放一个，最迟须在释放后 9 小时内完成。另有贯穿 48 小时的交互负载 U_t ，模型不允许把它移到其他时段，从而保留其即时性。每个交互工作单位的合成能耗设为 0.8 kWh，每个批处理单位设为 1.0 kWh；每小时另加 18 kWh 固定 IT 能耗，代表空闲 / 待机；每小时 IT 能耗上限为 150 kWh。本实验中的“工作单位”不直接对应词元数量、Groq 芯片或真实中心容量。它只是分析时序选择的抽象计量。

每种冷却方式的合成电力曲线，是时刻 t 的 IT 能耗比例 $c_{m,t}$ ，其中 $m \in \{dry, evap\}$ ；对应的用水曲线 $w_{m,t}$ 以 L/kWh IT 计。在复现代码中，干式冷却负荷随室外温度升高而增加；蒸发冷却在部分时段用电更低，但耗水更多。干式选项仍包含 0.035 L/kWh IT 的小额非零合成用水，用来表示辅助消耗，因此不能把它描述为绝对零水过程。

输入项	合成设定	纳入原因	不能据此推断的内容
时间范围	48 小时	允许在虚拟日内调度	不是完整一年的季节变化
最大容量	150 kWh IT / 小时	产生合理拥塞	不代表商业机架容量
待机用电	18 kWh IT / 小时	防止把迁移任务误写成消除全部能耗	不是 Groq 空闲功耗测量
交互单位用电	0.8 kWh	统一必须即时完成的工作	不等于一个词元单位
批处理单位用电	1.0 kWh	让工作迁移可计算	不等于 GPU / LPU 指标
批处理弹性	最多 9 小时	限制负载迁移	不能推广到所有 Batch API
λ	0 至 0.12 kWh/L	描绘数学权衡	不是水价或官方权重

生成这些输入的原始命令保存在 `analysis/reproduce.py`。保留代码而不只保留近似表格很重要，因为读者可以检查分段函数（`max`）、冷却切换点，并修改参数后重新运行。

8. 数学问题及其可解性

设 $x_{j,t} \geq 0$ 为批处理任务 j 在小时 t 完成的工作量。只有当 t 位于任务释放时间 r_j 与截止时间 d_j 之间时，该变量才允许存在。设任务所需工作量为 b_j ，则每个任务必须完整完成：

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

在每小时功率保持不变的模型假设下，逐时能耗约束为：

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

这里采用线性规划（linear programming, LP），因为批处理工作可以拆成更小单位。现实中可能存在不可拆分任务、KV 转移、模型实例启动以及只能取整数的批量，此时需要混合整数线性规划（mixed-integer linear programming, MILP）或在线分层求解。本文不声称分数解已经覆盖这些细节。

在建立线性规划成本前，程序先逐时选择 $m_t^* = \arg \min_{m \in \{dry, evap\}} [c_{m,t} + \lambda w_{m,t}]$ 。固定 m_t^* 后，该小时每个批处理工作单位的成本为 $1 + c_{m_t^*,t} + \lambda w_{m_t^*,t}$ ，优化目标为：

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

交互负载与待机负载仍计入所有策略的最终评估。在冷却方式固定时，它们相对于 x 为常数，因此没有出现在调度目标的可变部分。冷却方式不是这个 LP 内部的决策变量；这里采用的是先逐时取最小值、再优化批处理分配的可分两阶段求解。它只有在成本线性可分、各小时相互独立、冷却方式可以自由切换且不存在切换成本、最短运行时长或其他跨时约束时，才与联合优化等价。现实设施能否在一小时内完整切换干式或蒸发式冷却，也取决于实际设计：一座没有蒸发冷却设施的建筑，不能靠软件按钮增加冷却塔。

实验使用 SciPy HiGHS / linprog 求解器，为每项任务建立等式约束，为每小时建立容量约束。两个参考策略是在任务最早可运行时执行（FIFO），并分别始终使用蒸发冷却或干式冷却；另有两个补充策略：FIFO 加逐时最低电耗冷却，以及先逐时选择冷却方式、再在任务时窗内迁移任务的优化解。敏感性分析在 0 至 0.12 之间使用 8 个用水权重重复求解；公开数据包包含全部点，而不只是展示中选出的点。能耗函数没有 GPU 或 LPU 的硬件刻画，因此任何结果都不能读成芯片比较。

9. 运行结果与内部检查

复现脚本已使用 SciPy 在本地运行；运行说明、输入副本和结果均包含在复现包中。下表从每个天气情景中列出 7 种策略。所有结果覆盖相同时间、处理相同负载。FIFO-energy 表示任务一释放就执行，并在每小时选择电耗更低的冷却方式；LP λ 表示采用指定用水权重的线性规划解。电耗和用水量取决于本研究构造的函数，不能赋予模型之外的统计含义。

合成天气	策略	总用电 kWh	场内用水 L	干式冷却小时数
炎热干燥	FIFO—蒸发式	5223.07	6818.11	0
炎热干燥	FIFO—干式	5490.52	165.20	48
炎热干燥	FIFO—最低电耗	5192.71	4829.34	18
炎热干燥	LP, $\lambda=0$	5179.88	4166.87	18
炎热干燥	LP, $\lambda=0.02$	5188.14	3559.52	22
炎热干燥	LP, $\lambda=0.05$	5243.32	2222.32	30
炎热干燥	LP, $\lambda=0.12$	5407.56	165.20	48
沿海湿润	FIFO—蒸发式	5277.24	7076.90	0
沿海湿润	FIFO—干式	5504.52	165.20	48
沿海湿润	FIFO—最低电耗	5261.37	5451.87	14
沿海湿润	LP, $\lambda=0$	5245.08	4893.67	14
沿海湿润	LP, $\lambda=0.02$	5263.52	3640.80	22
沿海湿润	LP, $\lambda=0.05$	5313.90	2248.73	30
沿海湿润	LP, $\lambda=0.12$	5438.69	165.20	48

内部一致性检查。每项策略的合成工作量 W 都约为 4394.2 个任意单位，其中包括固定的即时工作量与批处理工作量；这不是任务数。另行检查的 48 个批处理任务均已完成，没有任务早于释放时间或晚于截止时间运行，也没有超过 IT 上限。代码核对总成本、拆分工作与所需工作量是否相等、每个工作片段是否位于任务时窗内，并拒绝求解器未成功返回的结果。这些断言只能检验模型是否按写定方式正确执行，不能检验热系数是否现实、客户如何行动，或 Groq 设备的真实行为。

读取表格时，一个重要比较是：炎热干燥情景从 LP $\lambda=0$ 转到 LP $\lambda=0.02$ 后，48 小时用电增加 8.26 kWh，场内用水减少 607.35 L；沿海湿润情景则增加 18.43 kWh，场内用水减少 1252.87 L。这些是合成程序的精确结果，由此计算出的百分比不能被表述为沙特计划的经济或环境回报。两个情景的比较只说明一个方法问题：不同的冷却函数，会改变“多用一千瓦时电能换取多少升节水”的权衡点。

两组合成天气数据上的水—电权衡前沿

图 1：每个点都来自一次实际求解，参数 λ 已明确标出。图中没有任何点来自现场测量或真实电网。

10. 为什么计算模型能成功，却仍不能代表机器

理论上，移动批处理任务像是无成本地交换时段；现实中，收益可能在四个环节失效。

第一，静态功耗与关机策略。如果机架空闲时仍消耗接近峰值的电，把任务移走也不会大幅减少能耗。若运营商采用深度关机，唤醒不会瞬间完成，模型加载和重新填充存储器可能需要数分钟。本模型加入固定空闲负载，避免把设施电耗简化为每项任务能耗之和；但它没有测试关机、唤醒或为突发即时负载保留容量的策略。

第二，模型在 SRAM/HBM 中的常驻状态。从一个时段移到另一个时段的批任务，可能碰到存储容量限制。包含大量权重与 KV 状态的模型需要跨芯片分片，卸载某个状态再恢复另一个状态的成本，可能吞掉制冷收益。现场实验必须把模型常驻（model residency）、KV cache 占用（KV cache occupancy）、交换流量（swap traffic）、编译缓存（compilation cache）和互连负载（interconnect load）与用电同步测量。

第三，服务要求与模型混部。单独看批处理拥塞似乎无关紧要，但同一硬件上可能还有交互用户。每一服务类别的 TTFT 尾延迟与请求完成时间都必须保持在时限内，不能只看平均值。上下文和输出长度差异，也会彻底改变预填充与解码负载。真实实验应抽样请求长度与到达波形的分布，不能把所有工作假设为均质单位。

第四，多回路热力学。冷却效率由换热面积、换热器类型、液体泵送成本、服务器入口温度、降雨与粉尘、补水处理等共同决定，不只取决于空气温度。如果没有制造商规格，就不能科学地假定一套设施每小时都能在干式和蒸发式模式之间切换。现场研究应包含真实的部分负荷性能（part-load performance）、冷却逼近温度（cooling approach temperature）、风机曲线、泵曲线和模式切换成本。

这里存在一条认识论边界：可行性证明（existence proof）不同于效应估计（effect estimate）。前者证明，在选择支持这种结果的参数时，时间弹性可以转化为多资源优化；后者还要证明参数代表真实机器、天气和负载，并在公平的任务分配与测量条件下与现有运营系统比较。

11. 知识上的先行者：此前已经有哪些工作

这个科学问题并非始于本文。2025 年的伯克利综述研究了服务器效率、发电用水、电网、冷却类型与气候变化之间的关系。美国能源部把 PUE 与 WUE 作为有用比率，同时提示计量边界限制。也就是说，区分设施地点与资源消耗，早在本文之前就是已有知识。[S11](#) [S13](#)

2026 年论文 *Balancing Bits and Drops: Stress-Adjusted Water Management for Data Centers* 研究了水压力的空间和时间差异，并提出把用水纳入工作负载安排的方法，也讨论了干式冷却与场外用水。它直接反驳任何“本文发明了首个用水调度器”的说法。本文建议验证的增量更窄：在沙特资源条件下研究多阶段推理，并分开处理任务类型、芯片、冷却与逐时数据。[S15](#)

同年 8 月发布的预印本 PowerSlider 研究瞬时功率约束，把预填充 / 解码的不对称性和服务质量时限纳入短周期迭代方案。作者报告，在其测试环境中，功率降低 30% 时，实验 Goodput 保持在 78.3%，而最强基线为 47.6%。这些数字不能转用于 LPU、Groq 或沙特；论文的电力目标也不等于本文的冷却用水目标。它仍证明，感知推理阶段的功率调度不只是理论讨论，而是已经有执行模型、负载与公开实验可供质疑和复验的研究方向。[S16](#)

供应商论文和公告也给出了可检验指标。NVIDIA 在 Groq 3 LPX 平台上公布 Gemma 4 31B 输出基准，并描述智能体工作负载中的高性能。公平处理这些结果，需要把它们限定在具体平台、模型、上下文和基准条件内，不能把它们改写成沙特全部服务的平均值。单用户每秒词元、总词元吞吐与完整任务时间并不相同，选择比较对象时必须使用同类指标。[S07](#)

12. 一项能在假设错误时推翻它的沙特现场实验

本节是尚未执行的提案，与上文已经运行的 Python 合成实验不同。它需要数据中心运营方授权、经确认的电表与水表读数，以及在不暴露用户数据、不危及生产系统的情况下复制负载的能力。研究无须公开专有架构，只要发布足以理解方法和复核方程的受控统计汇总。

样本与分层。收集数周真实运营轨迹，覆盖温湿度和请求长度变化。事先把工作分为三类：有严格时限的实时对话、允许合理完成时间的多轮智能体任务，以及具有真实截止期限的批处理。每项实验策略都保存输入与输出长度分布、时间戳和模型版本。私人文本应从研究数据中排除，或在明确可比内容后转成匿名长度指标。

比较组。（a）现行生产调度器；（b）只感知批任务时限与 GPU / LPU 负载的调度；（c）感知时限与电力的调度；（d）感知时限、电力与场内用水的调度；（e）如有可靠来源，再增加发电耗水因子的计算组。各组必须遵守相同的质量、服务与可用功率上限，并设置在比较开始伤害真实请求时立即停止的保护机制。最后一组没有预设优势；若它未能在置信区间内提高单位资源的有用工作量，其主张就应被拒绝。

硬件层。只有在同一模型、精度与质量具有可比性时，才比较通用 GPU、LPU 或混合平台。必须记录数值精度（precision）、量化类型、上下文长度、KV 容量、模型加载位置、并发用户数、网络 and 工具调用序列。不能把一款加速器上的网页演示与另一网络中的生产系统相比，再把全部差异归因于芯片物理。如果某个模型无法在特定芯片上以相同质量条件运行，这本身是重要的适用性结果，不应从报告中删除。

环境层。同步采集 T_{db} 、 T_{wb} 、冷却液供回温度、风机与泵功率、补水表读数，并把取水量与耗水量分开；如可获得，还应记录水源与水类型。IT、PDU、UPS 和冷却用电都应在明确的计量边界读取。研究需公开仪表精度范围、采样间隔，以及电力时钟与请求时钟之间的偏差，避免用延迟到达的测量制造虚假的逐时相关性。

成功指标。首要指标是按时且正确完成的任务数、每千项合格任务的设施能耗、每千项合格任务的场内耗水。PUE / WUE 只作为并列指标，同时报告 TTFT 以及 p50 / p95 / p99 完成时间。必须报告质量失败率和重试率，因为更快地产出错误答案不等于生产率更高。还应计入准备和编译多个模型所需的时间与

运营成本，不能把整日工作的可用范围压缩成某个瞬间的最快词元。

因果识别边界。不能简单比较一个气温温和的星期与一个炎热星期，因为请求组合、软件版本、天气、可用功率和能源价格都可能同时变化。更稳妥的做法是在条件相近的时窗内，用授权测试负载做随机交替；控制上一实验组留下的机架运行状态、模型预热等残留影响，或采用含清洗期的交叉设计（crossover design）。研究者还应预先公开判定结果所依据的分配方案，并披露负结果和条件差异，再提出全国层面或十亿美元项目层面的结论。

13. 下一阶段的经济：从处理器价值到把电转成有效时间的价值

拥有计算设施的城市可以出售账面容量、原始输出词元数，也可以出售在质量与时限约束下完成的任务。三者所承载的工业能力不同。第一种商品是已配备的容量；第二种是更细的计量产品；第三种则要求工程团队处理上下文长度、模型与处理器匹配、网络带宽、调度、能源与水表，以及确保工作真正有用的实验。即使不增加加速器，改善负载分配也可能提高产出价值；实现成本与复杂性也可能抵消收益。这不是收益承诺，而是一个可检验的微观经济假设。

在沙特，可以探索与 Groq 原生服务相邻的运营能力：开发异构架构间的推理负载转换工具、结合本地条件测量的调度算法，以及把模型类型、上下文和阿拉伯语等语言质量相连接的基准库。这类知识可以进入测量“每焦耳合格任务数”和“每升场内用水合格任务数”的平台，并区分购买几分之一秒的即时任务与拥有数小时弹性的任务。但不能仅凭全球市场投入巨大，就推断产品必然有销量或防御性优势。软件产生的决策价值，必须与 NVIDIA、Groq 的原生方案、大学研究和冷却供应商方案比较验证；否则只是给已有工具换名。

2027—2035 年的方向不是数值预测，而是一棵测试树：如果可用的标准化冷却数据出现，平台也能以有限传输成本管理多种处理器，物理感知调度或许值得进入推理服务的控制平面（control plane）；如果绝大多数负载仍必须即时处理，或状态迁移消耗超过时序优化的节省，则应优先通过编译器、量化和电气效率降低每词元能耗。如果场内节水与发电端用水相冲突，在明确核算边界前就不能宣布获胜。这些结论都取决于实验结果，可以修改，不是关于 2035 年必然发生什么的断言。

14. 计算复现与独立验证

复现包包含代码、输入数据和可供审查的结果：

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

代码依赖 NumPy、SciPy 与 Matplotlib，不需要 API 密钥，也不调用任何专有产品内部接口。

data/weather_synthetic.csv 与 data/workload_synthetic.csv 是合成输入；

data/policy_hourly.csv、data/policy_summary.csv 和 data/policy_sensitivity.csv 是详细结果；

data/run_manifest.json 保存运行设置。evidence_workbook.xlsx 提供可检查的结果表、计算公式和输入引用。图表同时保存为 PNG 与 SVG，figure_register.csv 明确标记其合成性质。完整复现材料见 S17。

本文没有来自气候或企业的概率样本，也无法为产业效应提供有理据的置信区间分析。HiGHS 求解器解决的是一个明确写出的线性问题；它在给定假设下正确求解，不会自动把结果变成行为估计或全国估计。我们没有测试沙特数据中心的真实队列，也没有确认 Groq 芯片的每词元能耗。复核运行能够把表中汇总结果重现到所示精度；但在线性目标存在并列最优解时，同样最优的逐时调度结果可能不同。一次复核中，沿海情景的 1056 条逐时记录有 48 条出现这种差异。因此，若不固定软件环境并加入确定性平局消解规则，逐时 CSV 与由其生成的图形未必逐字节相同。读者可以重跑代码，再用经授权的实测关系替换合成曲线，检验结果如何变化。必须区分计算可复现性（computational reproducibility）与独立实证复验（empirical replication）：前者由复现包支持，后者仍是尚未开展的科学项目。

15. 目前未知、图表也不应假装知道的事

现有公开来源没有披露 Groq 达曼中心的电耗、水耗、排热设计，各地点的芯片类型与数量，沙特境内单独生成的词元量，或与 15 亿美元投资承诺相对应的新增水资源容量。本文没有找到证据把 NVIDIA Groq 3 LPX 新平台与任何具体沙特中心相连，尽管 Groq 宣布计划成为早期采用者之一。我们也没有在统一任务条件下，从沙特境内独立证明 Groq LPU 与 GPU 的性能差异。“HUMAIN 推理服务提供商”是一项公布的商业关系，不是真实账单记录，因此不能据此推导使用量或收入。

本文也没有取得 NCM 的逐时数据；实验中的“沿海”只是合成天气弧线。没有测量任何中心的 WUE，也没有计算海水淡化边际用水或电网发电用水。真实中心可能采用不同的冷却方式，或者不允许模型假设的切换。这些限制不是靠增加建议性文字就能修补的表述缺陷，而是现有证据能够支持的精确边界。

可质疑的结论：沙特已经出现 Groq 推理服务的真实运营信号，也存在测试一种软件的机会：它是否能提高单位电力和水资源下，按时完成的正确工作量。但科学与经济上的成功，不能从“LPU 在所有方面都更好”的假设出发，也不能把切换到干式冷却直接等同于可持续性。它应从两个运营问题出发：任务速度真正受什么限制——存储、计算、网络还是队列？这项任务在这个地点、这个小时，按可审计计量表究竟用了多少电和水？

参考资料与追溯来源

以下条目优先采用机构原始材料或学术论文。供应商新闻稿只能证明其自身发布过相关陈述，不能代替独立验证。

1. [S01 — Groq / Aramco Digital: 沙特推理数据中心合作公告, 2024-09-12](#)
2. [S02 — Groq: 达曼运营与 15 亿美元扩建承诺, 2025-02-10](#)
3. [S03 — Groq: 成为 HUMAIN 推理服务提供商, 2025-05-27](#)
4. [S04 — Groq: LPU 架构说明, 2025-03-07](#)

5. [S05 — GroqChip Processor Product Brief v1.7, 2024](#)
6. [S06 — Groq: 与 NVIDIA 达成非独家推理技术许可协议, 2025-12-24](#)
7. [S07 — NVIDIA: Groq 3 LPX 全面投产, 2026-08-24](#)
8. [S08 — Groq: 平台和基础设施网络页面](#)
9. [S09 — GroqDocs: 服务层级与批处理接口](#)
10. [S10 — GroqDocs: Prometheus 请求延迟指标](#)
11. [S11 — 美国能源部: 联邦数据中心冷却用水效率, 2019-01-09](#)
12. [S12 — Groq: 规划采用 NVIDIA Groq 3 LPX 与 Vera Rubin, 2026](#)
13. [S13 — Lei 等: *The water use of data center workloads*, 2025, DOI 10.1016/j.resconrec.2025.108310](#)
14. [S14 — 美国国家标准局 / NIST: 约 25° C 时水的汽化潜热测量](#)
15. [S15 — Talukder 等: *Balancing Bits and Drops*, arXiv:2607.22617, 2026](#)
16. [S16 — Li 等: *PowerSlider: Exploiting Phase Asymmetry*, arXiv:2608.21719, 2026](#)
17. S17 — 48 小时水—电—工作线性规划复现包; 输入完全合成
18. [S18 — 沙特国家气象中心: 天气 API 入门文档](#)