



战略文章

编辑日期：2026 年 10 月 23 日

# 当达曼开始制造“秒”

Groq、专用推理芯片与有用人工智能的热力学

在一次面向沙特观众的演示中，真正值得注意的，不是语言模型回答了一个与“2030 愿景”有关的问题——许多模型都能做到这一点。值得注意的是，Groq 创始人 Jonathan Ross 与时任 Aramco Digital 负责人 Tareq Amin 在 2024 年全球人工智能峰会上试图展示的体验：人与机器进行语音对话时，不再因为等待回答而不断打断自己的节奏，计算似乎不再是显眼的障碍。这是两家公司安排的演示，不是一项公开发表、经独立验证的延迟测试；但它把 Groq 的核心问题摆到了台前：如果两个词之间的等待，本身就成了处理器设计问题，会发生什么？[S01](#)

几个月后，一则不同于常规数据中心建设的消息出现了。Groq 称，它在 2024 年 12 月用 8 天时间让一个沙特推理集群投入服务；2025 年 2 月，公司又称其达曼中心已通过 GroqCloud 处理全球客户请求，同时公布了用于扩建基础设施的 15 亿美元沙特投资承诺。这些数字和事实均来自公司自己的公告。它们不表示全部资金已经支出，不表示所有规划容量已经落地，也不是一份推理服务出口账本。不过，它们足以改变沙特在技术地图上的位置：抵达沙特手机的回答未必都在境外计算；某个海外开发者发出的请求，也可能在沙特境内完成计算。[S02](#)

由此引出的议题，比服务器贸易更深一层。为什么保存模型权重的地点，有时不如能在几分之一秒内吐出下一个词元的地点重要？为什么一家为生成语言而造芯片的公司，会把海湾地区设为全球网络的一站，后来又与看似天然竞争对手的 NVIDIA 建立技术许可和架构协作关系？还有一个会打破漂亮数字的问题：如果沙特生产了更短的等待时间，耗掉的电去了哪里，又是哪一部分水把热量带走？

本文讨论的不是服务中断，也不是治理，更不把“模型主权”当作口号。

本文讨论的是推理时间的工业化生产。它从芯片内一小块存储器的电学活动开始，经过编译器（compiler）、调度器和热管，最后落在一张可以核算的资源账单上。这里存在一个并不直观的沙特机会：出口可测量的时间性能，而不只是托管一个被称为模型的文件。

## 回答并不是一次性说出来的

假设你向语音助手询问一份工业报告。听者听到的是连续句子，推理系统看到的却是神经网络按概率依次选择的一串词元（token）。问题先进入预填充阶段（prefill）：上下文词元相对并行地处理，系统建立内部表征以及注意力的键值缓存（key-value cache, KV cache）。随后进入解码阶段（decode）：模型生成一个词元，再把它并入生成下一个词元所需的上下文。这不是文学性的时间切分，而是两类计算物理过程。

预填充阶段集中了矩阵乘法 and 大规模计算；单个请求的解码阶段则反复读取模型权重和注意力状态，限制因素可能更多来自存储器数据搬运，而不是算术运算。因此，仅说一块芯片每秒能做更多万亿次运算并不够。更精确的问题是：每从存储器取回一个字节，能完成多少计算？性能工程师把它称为算术强度（arithmetic intensity），并用屋顶线模型（Roofline model）判断程序何时受计算能力限制、何时受存储器带宽限制。

工厂增加生产线可以提高总产量，却没有回答眼前顾客要等多久。同样，总吞吐量（aggregate throughput）不同于单用户每秒词元数（tokens per second per user）。服务商可能因为把数千个请求合在一起而得到惊人的总数，单个用户却仍觉得模型说话缓慢。两者之间至少有四项不同指标：首词元时间（time to first token, TTFT）、词元间延迟（inter-token latency）、排队延迟，以及从发出请求到最后一个词元返回的端到端延迟（end-to-end latency）。Groq 的运维文档分别列出了这些可观测指标，也区分了具有严格响应要求的服务与无需即时交互的批处理工作。[S09](#) [S10](#)

用一个与特定芯片无关的算例说明：生成 180 个词元，速度为每秒 300 个，理论解码时间是 0.6 秒；速度提高到每秒 900 个后，时间变成 0.2 秒，只在这个阶段节省 0.4 秒。它没有计算把语音传给模型、语音转换、外部工具读取，也没有计算生成前的等待。但如果一个软件智能体要连续进行四轮同样规模的生成，这一阶段累计可节省 1.6 秒。在智能体系统（agentic systems）中，串行词元速度的影响会随步骤数放大，因此推理的经济学不同于一次性训练模型的经济学。

## 为什么 Groq 把调度带进芯片

Groq 将其架构称为语言处理单元（Language Processing Unit, LPU）。这个名称更像公司对一类专用计算的商业定义，而不是计算机体系结构教科书里公认的门类；名称背后的工程选择仍值得拆解。现代图形处理器（graphics processing unit, GPU）功能强大且用途广泛，依靠高带宽存储器（high-bandwidth memory, HBM）、Tensor Cores、调度器、指令集和软件层处理多种深度学习负载。Groq 采取的另一条路线，是在执行开始前，就让数据与操作的流动顺序大体确定。[S04](#) [S05](#)

传统芯片上的程序，会在运行中询问哪些资源已可执行下一步。另一种方式是静态调度数据流

（statically scheduled dataflow）：编译器提前把操作和数据移动排入时间表，结果在处理单元、存储器与互连之间移动，仿佛一条提前知道材料何时抵达每个工位的生产线。这是对公司所述架构的解释，并不表示外部服务条件也全部变成确定性的。队列、网络、用户争用和模型差异仍会造成波动。不过，芯片内部的数据移动更可预测，意味着硬件在运行时需要临场解决的问题更少。[S04](#)

Groq 还选择了片上静态随机存取存储器（static random-access memory, SRAM），而不是主要依赖从芯片外的 HBM 取回权重。SRAM 用快速晶体管单元保存比特，不需要动态随机存取存储器（dynamic random-access memory, DRAM）那样的刷新周期；代价是每个比特占用更多硅面积。它用存储密度换取速度与低延迟。这种取舍才是芯片设计的关键，而不是一句“存储器快十倍”。Groq 公布过某一代芯片 SRAM 带宽超过 80 TB/s，并将它与部分 GPU/HBM 配置约 8 TB/s 的数据对照。这是受数据路径和产品代际定义约束的供应商比较，不能据此评价所有 GPU，也不能据此断言任何模型的完整推理都快十倍。[S04](#)

真正棘手的问题看似简单：权重放在哪里？一个假设的 700 亿参数模型采用 16 位精度时，仅权重就需要约 1400 亿字节，即 140 GB（十进制），尚未包括注意力状态、临时空间、冗余和芯片间协调。若权重采用 8 位表示，原始体积降为 70 GB，但部署并不会因此变成零成本：数值精度、量化尺度、KV cache 和分片方式仍需处理。NVIDIA 在 2026 年公布的 Groq 3 LPX 机架设计包含 256 个加速器，合计 128 GB SRAM，公布的机架内存带宽为 40 PB/s。因此，即使在较新一代平台上，一个 70B 模型的 16 位原始权重也可能超过单机架 SRAM 总量。这不表示系统无法提供服务，因为还可以采用权重分片、量化，以及其他存储器、加速器与组件。[S08](#)

由此可以推断：数据越靠近计算单元，读取等待通常越短。同时，软件也必须被视为一张跨芯片的权重、存储和通信分布图。可以在沙特本地发展的能力，不止是采购芯片，还包括编译器感知的放置（compiler-aware placement）、不破坏应用质量的量化、变长批处理调度，以及测量 KV 数据在资源间移动的频率。建立这些运维能力，并不要求先在短期内自行发明一块芯片。

GPU 与 LPU 存储流的概念比较

图 1：解释性原创示意图，不是制造原理图，也不是两类芯片的性能测量。原理来源：[S04](#)、[S07](#)、[S08](#)。

## 三次真实事件，三种不同的“成功”

第一，从演示台到达曼。2024 年 9 月，Aramco Digital 与 Groq 宣布合作建设大型推理中心，并设想与数字服务市场 *nawat* 相连。台上出现了一段语音对话。它证明了一个交互界面的演示场景，不能独立证明全部规划基础设施已经建成。随后，供应商宣布 12 月用八天启动集群，又在 2025 年 2 月宣布从达曼中心提供 GroqCloud。三者性质不同：合作签署创造了实现的可能，演示证明一个窄场景可运行，服务上线公告则表示面向开发者的产品据称已进入运营。现有公开资料没有给出达曼中心逐时电耗、水耗或独立词元量。[S01](#) [S02](#)

第二，本地模型无需等到芯片也在本地制造，才可以在本地运行。Groq 在 LEAP 2025 展示了沙特 ALLaM 模型，以及英语和阿拉伯语推理、语音生成模型。2025 年 5 月，公司公告称自己成为沙特 HUMAIN 的推理服务提供商。前者展示模型和用例，后者公布供应关系。这里需要保持三个工业层次的区别：模型可以在沙特开发，处理器可按 Groq 的披露在美国设计和制造，而请求的计算地点可以位于沙特。它们是价值链的三层，不会因为都被冠以国家名称而合成一层。[S02](#) [S03](#)

第三，竞争者进入了同一台机器。2025 年 12 月，Groq 宣布向 NVIDIA 非独家许可推理技术；Jonathan Ross、Sunny Madra 和一批工程师转入 NVIDIA，而 Groq 按公告仍是独立公司。2026 年 8 月，NVIDIA 宣布 Groq 3 LPX 全面投产，Nebius 是首家采用者；Groq 则表示，公司计划与 Dell 合作，成为首批部署者之一。“产品已投产”不能被读成“已经装进沙特中心”，因为现有材料没有提供这种证明。[S06](#) [S07](#) [S12](#)

更值得注意的是，NVIDIA 的新设计把两类计算放到一起：基于 GPU 的 Vera Rubin NVL72 与专用 LPX 机架。据 NVIDIA 描述，两者共同计算模型的不同层，以服务长上下文和高速生成。这不是一块芯片获胜、另一块退出历史舞台的故事；它反而承认：推理并非必须由同一种硅完成的单一操作。密集的上下文处理可以交给一种计算形态，串行词元生成可以交给另一种，具体分工取决于工作负载、网络与存储。NVIDIA 引述 Artificial Analysis，公布了 *Gemma 4 31B* 在 10 万词元上下文测试中约每秒 3400 个输出词元的基准结果。这个数字只对应特定模型、测试和平台，不是对所有客户的承诺，也不是达曼中心的测试结果。[S07](#) [S08](#)

地理信息还有一个小矛盾，足以阻止我们匆忙画图。2025 年的公告仍写明运营地点是达曼；Groq 当前平台页面则在网络中列出 SA-1 — Riyadh（利雅得），并称全球共有 13 个中心。页面没有解释这究竟是不同的运营标识、新增地点，还是接入点分类方式变化。不能把达曼和利雅得直接相加成两个独立中心，也不能假定业务已从一地迁到另一地。可以确认的是，沙特已成为 Groq 公布网络中的运营节点；无法确认的是其准确构成与当前负载分布。[S02](#) [S08](#)

## 芯片吐出的词元，未必在有用的时间抵达

处理器即使每秒生成 1000 个词元，若先在队列中等待很久，应用仍会失去规格表上看不见的价值。如果智能体要跨洲调用四个工具，加快解码也不会消除网络传输和工具执行时间。因此，请求时间图谱比单一速率更有用：网络往返时间（round-trip time）、排队时间、预填充时间、解码词元间隔、工具等待，以及整个任务在达到目标质量时的完成率。

Groq 的商业文档已经给出一个无需猜测其内部实现的现实切面：有面向严格响应要求的 performance 层、面向常规请求的 on\_demand 层，以及以部分服务保障换取更宽处理边界的 flex 层；另外还有时限和边界不同的 Batch API。这不证明公司已运行热感知调度器，却证明现实推理服务中的工作具有不同的时间价值。实时电话回答不能等待更凉爽的天气窗口；大型档案摘要或一批模型评测则可能有数小时的弹性，前提是合同和数据性质允许。[S09](#) [S10](#)

因此，沙特云服务运营者可以定义一种更严格的商品：在明确时限内交付、质量合格的词元，并核算生产它所需的电与水。原始词元数会误导：系统可能生成无人需要的冗长内容，重做失败答案，或在期限之后完成批次。芯片空闲时的耗电，以及制冷系统再次消耗的能量，也根本不会出现在词元计数器上。产业需要三个相互交叉的计量表：有用工作、有效时间和物理资源。

## 天气不是建筑背景，它可能成为算法输入

在计算机科学中，把任务放进合适时间窗称为调度（scheduling）。如果把室外空气也纳入任务的成本函数，会怎样？手机天气应用显示的温度不够。制冷工程师区分干球温度（dry-bulb temperature）与湿球温度（wet-bulb temperature）：前者是普通温度计显示的空气温度，后者反映在现有湿度下通过蒸发降温的潜力。炎热干燥的空气与沿海湿润空气，适合的蒸发冷却方式不同；从空气还能接收多少水汽来看，沿海夜晚也未必等于“凉夜”。

推理中心里的电能不会消失，最终大多变成热，必须从处理器、存储器和电源系统带走。液体可以流经冷板（cold plate），再通过换热器和泵把热量送至设施外部的排热系统。液体在闭式回路中搬运热量，不等于设施绝不耗水；同样，使用液冷也不等于一定配置蒸发冷却塔。水的影响取决于最终排热方式、补充水、处理和排放，以及电力来源。[S11](#) [S13](#)

干式冷却（dry cooling）通过机械方式把热排入空气。它的场内直接蒸发用水很少，但在高温时可能消耗更多电。蒸发冷却（evaporative cooling）在某些条件下可降低制冷用电。它同时消耗场内用水。不存在对所有地点、所有时段都占优的方法。美国能源部资料将电能使用效率（power usage effectiveness, PUE）定义为设施总能耗与信息技术设备能耗之比，将场内水资源使用效率（water usage effectiveness, WUE）定义为设施用水量与 IT 用电量之比。但这两个指标本身都没有告诉我们完成了多少项有用任务，也没有包含设施边界之外发电过程的耗水。[S11](#)

一个小计算可以说明其中的物理量级：在约 25° C 时，蒸发 1 千克水需要约 2.44 MJ 潜热；1 kWh 等于 3.6 MJ。若作一个纯粹理想化的假设，把每 1 kWh 热量都完全通过蒸发移除，理论上会蒸发约 1.47 L 水。这不是任何数据中心的 WUE，不是每个问题的用水量，也不包括排污、损耗和电力供应链用水。它只是能量与物质相变的关系，提醒我们：比特看似没有重量，它产生的热却有物理代价。[S14](#)

## 可复现实验：调度器能否把热量从一个小时移到另一个小时

我们为本文运行了一个小型计算实验：执行过程是真实的，输入完全是合成的。实验设置一个 48 小时时窗，构造两条虚拟天气序列，一条“炎热干燥”，一条“沿海湿润”。我们没有把它们命名为利雅得或达曼，因为没有取得沙特逐时观测数据；它们只是形状合理的数学情景，不是天气记录。每小时都有必须立即执行的交互负载，也有最多可延迟 9 小时的弹性批处理任务。模型还设定 IT 功率上限、固定空闲功耗，以及干式与蒸发式冷却的电力和用水函数。在成本线性可分、冷却方式可逐时自由切换且没有切换成本或最短运行时长的假设下，程序先逐时选择目标成本最低的冷却方式，再由线性规划（linear programming）在保留全部截止期限与相同有用工作量的条件下选择批处理运行时段。目标函数里用于权衡 1 L 水与 1 kWh 电的系数，只是数学权重，不是沙特市场水价。研究伴随稿和复现包列出了全部假设、函数与方程。[S17](#)

在炎热干燥情景中，基线策略让批处理任务一有机会就开始，并在每小时选择用电更少的冷却方式；48 小时总耗电为 5192.71 kWh，场内用水为 4829 L。只以电力为目标时，程序先逐时选择最低目标成本的冷却方式，再以线性规划调整批处理时机，总电耗降至 5179.88 kWh，用水约 4167 L。当目标函数加入 0.02

kWh/L 的有限用水权重时，解变成 5188.14 kWh 与 3559.52 L。与最低电耗解相比，这相当于多用约 8.26 kWh，换取约 607 L 的用水减少。它不是可推广的真实节水比例，而是为本研究构造的冷却函数所产生的条件性结果。S17

沿海湿润情景的取舍不同。最低电耗解需要 5245.08 kWh 与 4893.67 L；采用同一用水权重后，结果为 5263.52 kWh 与 3640.80 L。在这个合成世界里，多用约 18.43 kWh，场内用水减少约 1253 L。两个情景中，48 个批处理任务全部在各自时间窗内完成；按共同的任意合成单位计，交互负载与批处理负载之和  $W = \sum_t (U_t + \text{batch}_t)$  均约为 4394.2。这个数是连续工作量，不是任务数，也不是概念性指标  $Y$  在  $v_j = 1$  时的取值。由此显示的是一种可能机制，而不是简单的“越冷越好”：关键在于把任务完成时间与热量和用水的边际成本相匹配。S17

两个合成情景中的制冷用水—用电权衡曲线

图 2：对合成输入运行线性规划得到的实际求解结果；每一点处理相同工作量，且没有超过截止期限。两轴使用的是合成模型单位，不是沙特设施测量值。

交互任务在实验中完全不能移动，因此这条曲线不是以减慢实时对话换来的。只有明确拥有时间窗的任务可以移动。热调度器不是所有请求的通用开关，它能做多少，取决于实际存在多少时间弹性。如果没有这部分空间，软件也不应假装自己能改变天气。

合成一天中批处理工作量的时间移动

图 3：炎热干燥情景中的批处理时序示例。曲线来自本地合成实验，不代表 Groq 或 HUMAIN 的实际负载。

## 从编译器调度器到城市调度器

Groq 在芯片内的工作与本文在设施层提出的设想之间，有一条有启发性的联系；它们共享协调逻辑，却承担不同功能。芯片层的编译器安排矩阵计算结果在各单元间移动，无法为了等待整座城市降温而暂停为实时用户生成下一个词元。服务层调度器则决定某批评测或文档处理任务，应放在哪个小时、哪个地点、哪种冷却配置上。前者以纳秒或时钟周期计，后者以分钟或小时计。两者面对同一个问题的不同版本：哪些工作可以移动，而不破坏正确性和时限？

严谨的工程表述必须避免夸大。其他研究已经在探索按用水量与地方水压力进行负载调度，包括 2026 年的预印本 *Balancing Bits and Drops*。同年 8 月，研究者在 SGLang 系统上发表了针对瞬时功率限制的 *PowerSlider*，说明降低 GPU 频率对预填充和解码的影响并不对称，并报告了特定实验配置下的 Goodput；该环境不是沙特数据中心，也不是 LPU。本文建议检验的增量，不是从零“发明环境调度”，而是在沙特设施中，把随时间变化的冷却条件与真实推理负载特征相连，再用可比测量评估达到质量要求且按时完成的工作。S15 S16

如果从实际运营团队的角度出发，可以先采集一条请求轨迹：预填充大小、输出词元数、KV cache 占用、工具调用步骤数、请求开始时间、首词元时间、用户可接受的上限，以及任务重试次数；再从物理层加入机架 PDU 用电、泵与风机状态、换热器效率、干球与湿球温度、补水表读数和冷却流量边界。精细量化或共享提示词前缀缓存（prefix caching），可能比把任务移到晚间更重要；将预填充与解码分配给不同专用资源，也可能优先于任何热策略。优先级不能由修辞决定，必须由单位资源产生多少有效词元的测量来决定。

## 这个设想可能错在哪里

谨慎的工程师可能指出，把批任务移到另一小时会堆积负载、抬高峰值或降低服务器利用率；如果芯片在空闲时仍持续耗电，节省量就会低于预期。另一个团队可能发现，把 KV 状态移走或把模型重新加载到较空闲芯片所耗的电与时间，超过了制冷收益。制冷性能也可能更多取决于供液温度，而不是室外气温。如果最低场内电耗的运行方式，对发电端用水反而更糟，仅看本地 WUE 就不够。

还有一个更强的经济反驳：如果 Groq 等平台让推理更快，应用可能增加词元消费，而不是降低总账单。智能体响应越快，用户可能让它进行更多轮阅读、试验和核验。因此，每词元能耗可能下降，总能耗却上升，甚至工作负载的构成本身也会改变。劳伦斯伯克利国家实验室 2025 年的研究显示，在不同假设和设施条件下，归因到数据中心工作负载的用水量可相差一万倍以上，驱动因素包括服务器效率、电网水强度、利用率、冷却系统和气候。这不是沙特的实测比例，而是对“每个问题固定若干升水”这类说法的科学警告。[S13](#)

最后，如果用户在另一大洲等待实时语音回答，网络距离可能比远端设施的制冷节省更重要。能够向全球出口计算的地点，不会自动成为每个全球请求的最佳地点。有些应用对网络延迟敏感，有些则允许异步提交并在之后取回批处理结果。成熟的服务需要两种不同的商业负载：即时负载购买几分之一秒的时间，弹性负载在截止期限内购买成本更低的能量。前者不应承担只有后者才能在不伤害体验的情况下测试的全部水资源约束。

## 2027—2035 年，沙特可以建设什么

在冠以任何全球称号之前，可以先在沙特推理中心做一项工业实验。选择三类带有明确服务合同的真实负载：实时语音对话、多步骤软件助手，以及截止期限为 8 小时的一批模型评测。在模型、许可和可用性允许时，分别在 GPU 与 LPU 平台上运行，并保持相同质量合格率、相同有效词元数和相同请求长度分布。测量 p95 首词元时间、各分位词元间延迟、每小时有用工作、机架与设施用电、场内用水；如有可追溯数据，再计入发电用水。实验应跨季节重复，而不是只在宣传活动当天运行。

这里没有主张在三年内凭空制造一块沙特 LPU。Groq 的原始披露明确说，其芯片在北美设计、制造和组装。更有价值的本地问题是：沙特团队能否在异构推理编译（heterogeneous inference compilation）、上下文与 KV cache 分配，以及按每项按时完成且正确的任务审计用电和用水方面形成领先能力？能否开发根据预计输出长度、词元跨网传输成本和天气条件选择地点与平台的软件，再像出口推理服务一样出口这套软件与运营知识？这是研发课题和运营产品方向，还不是已经兑现的答案。

更小的起点是：在一座中心内建立同步测量系统，选择少量弹性工作做 A/B 试验，让调度器读取天气条件。如果它能降低用水，同时不损害客户响应，也不把负担转移到水强度更高的电力上，就出现了真实收益。如果不能，负结果同样有价值，因为它能指出系统中真正起约束作用的边界。成功必须表现为真实的资源与工作价值，而不是图表上漂亮的移动。

接下来才是战略问题。上个世纪，一些产业以铝、金属和化肥的形式出口能源，即用大量电力重新塑造物质。未来几十年，另一类产业可能把电、硅和制冷转化为可以定价的认知时间：语音交互减少若干毫秒等待，或者科学智能体在项目期限前多完成数千次推理尝试。这是一个待检验的经济类比，不是两类产业之间的等值计算。两者的竞争核心都在于把资源转成可销售产出的效率；废热与等待存储器，也就不再只是设备房里的数字，而成为商品本身的缺陷。

回到开头的演示，Groq 展示的是一段似乎比平常更快的对话。它留下的战略问题不只关乎机器说话有多快：如果沙特能够为世界生产更少等待的时间，并能按任务核算这些时间消耗的电和水，会怎样？这样一来，数据中心就成为具有明确产量的工程设施，芯片只是其中一种工具，调度器读取天气，就像读取句子长度一样。闪亮的新机架照片无法回答这个问题。答案必须来自可复核的实验：从词元追踪到完成的多轮任务，再从任务追踪到计量表，最终返回一个其他工程师能够重新计算的结果。

## 编辑说明与来源

本文区分供应商自述、公开运营公告、合成计算和待开展的现场研究。Groq 与 NVIDIA 关于性能和扩张的陈述均按供应商原话归属，除非另有独立验证；达曼与利雅得的地点不确定性已明确说明。48 小时实验的所有数字来自配套合成代码，不是敏感运营信息，也不是沙特实测。汇总结果可复现至文中精度；由于线性规划存在并列最优解，在未固定软件环境且未加入确定性平局消解规则时，个别逐时调度结果及其图形可能不同。

## 参考资料

以下条目优先采用机构原始材料或论文。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