



战略文章

编辑日期：2026 年 10 月 24 日

一台主动压低自己工价的计算机

为什么智能服务降价，可能早于硬件回本？

The Token Deflation Trap: When Faster Inference Undermines Compute Payback

开篇：一颗处理器，完全可能在完成既定任务时表现出色，却在同一天落后于经济回报要求。这不是芯片失灵的谜题，而是一个行业的现实：它增加智能产出的速度，可能快过它为这些产出收费的能力。我们将先翻开一家美国公司的财务文件，再看一种模拟内存分页的算法，最后回到已经在沙特投入运行的算力。决定胜负的，也许不是纸面上能完成多少次计算，而是计算成果能否及时变得有用、可售。

1. 在会计账上，芯片能用六年；在市场上，每个季度都有新对手

CoreWeave 的 2025 年年报中有一句很容易被略过的话：自 2023 年 1 月起，公司把数据中心计算设备的预计使用寿命从五年调整为六年。披露文件称，调整依据包括硬件进步、软件改进和数据中心设计。这里的六年是预计会计使用寿命（Accounting Useful Life），用于分摊折旧（Depreciation）；它并不承诺设备能连续六年保持行业最高盈利能力 [S01](#)。

2026 年 1 月 5 日，英伟达（NVIDIA）发布 Rubin 平台，称其一体化设计在特定工作负载下，可能把推理令牌成本（Inference Token Cost）降至 Blackwell 平台的十分之一。这是供应商对目标技术能力的说明，不是实际成交价格披露，也不能代表所有模型与应用 [S03](#)。

随后是 2026 年上半年财务披露。CoreWeave 表示，截至 6 月末，剩余履约义务（Remaining Performance Obligations; RPO）达到 1037 亿美元。数字很大，定义更关键：这是尚未履行服务所对应的合同金额，受会计规则以及延迟、折扣等因素影响；它既不是账上现金，也不是已经确定的净利润。公司预计在之后 24 个月确认其中约 41%，再在接下来的 24 个月确认约 39% [S02](#)。

谈人工智能算力，人们习惯从芯片名称和显存容量说起。但当同一份档案里同时出现“六年”“令牌成本降低十倍”和“1037 亿美元未来服务”，真正有意思也更危险的问题就浮现出来：哪一种时间必须赢得这场赛跑？是折旧周期、新一代技术出现的周期、合同保护现金流的周期，还是客户愿意支付的价格发生变化的周期？

这些数字没有证明市场将要崩溃，也不能证明某家公司已经掉进陷阱。恰恰相反，长期合同的存在提醒我们，不要讲成“新一代发布，旧处理器立即归零”的故事。它们真正说明的是，投资者不该只问“芯片能不能工作”，而应问：“谁会购买它的工作时长，以什么价格，买多少次，又能否赶在资本到期以前？”

2. 开发者眼中的降价消息，到了融资方那里就是一条新假设

2026 年 9 月，OpenAI 宣布 GPT-6 Sol 和 GPT-6 Luna 的应用程序编程接口（Application Programming Interfaces; APIs）价格低于上一代对应产品的推广价格，并把部分降幅归因于缓存和推理改进。在公开的 Sol 输入价格示例中，每百万令牌从 4 美元降至 2 美元；输出价格则从 20 美元降至 10 美元。这是两代不同产品在公开定价条件下的比较，不能证明每一颗芯片的市场租价都减半 S04。

小型应用的开发者看到这条消息，可能会得到过去被推理账单压缩掉的试验空间：把五种写法扩展到十种，或者为原本负担不起更强模型的用户提供服务。买下整排处理器、又把采购款变成长年融资义务的人，则会问另一个问题：服务价格下降后，需求会不会以同样速度翻倍？还是说，为获得同样收入，必须卖出更多工作量？

我把由此产生的风险称为令牌价格通缩陷阱（Token Deflation Trap）。价格下降本身并不会触发陷阱。三个条件同时出现时，陷阱才会形成：每个售出工作单位带来的收入下降；有用且付费的工作单位无法按所需幅度增长；资本义务的成本又没有同步下降。只要其中一项不成立，陷阱就可能转为难得的机会。因此，这个名字不是预言，而是一道抓住关键环节的问题。

真实市场也远比“每百万令牌多少钱”复杂。输入模型的令牌和模型输出的令牌不是同一回事；复用已有上下文的缓存输入令牌（Cached Input Token）不是一次全新的计算；一句简短而有效的回答，也可能省去一长串重试。如果把“一百万令牌”当成一百万颗完全相同的钉子出售，很快就会发现：有些钉子占用显存，有些钉子需要等待，有些钉子原本就从缓存盒里取出。会计单位很整齐，物理单位却不听话。

3. 聊天窗口看不见的令牌旅程

要理解软件为什么能给硬件重新定价，不妨从一句普通用户的请求开始：“请概括这些邮件，并把财务决定和上周的往来联系起来。”图形处理器（Graphics Processing Unit; GPU）收到的不是一句已经准备好供其“思考”的话，而是一串令牌（Tokens）。它们将经过多种计算过程，在内存、传输与并行处理上的行为并不相同。

在上下文预填充阶段（Prefill），输入令牌先被处理，以建立注意力表示（Attention）。随后进入反复生成的解码阶段（Decode）：生成一个令牌，再把它纳入下一个令牌的上下文。预填充的计算密度往往更高，也更适合并行；具体程度取决于负载规模和架构。解码则可能受到内存带宽（Memory

Bandwidth)、权重搬运和对话状态的限制。一颗芯片即使擅长前一阶段,也不一定在下一阶段拥有最低的最终成本。TensorRT-LLM 等工具因此记录了分离两阶段资源的架构:需要时分别配置,而不是强迫它们采用同一种运行节奏 [S07](#)。

对话还有一种不同于长期记忆的工作内存。键值缓存 (Key-Value Cache; KV Cache) 保存可复用的表示,避免反复计算部分上下文。每一段长对话都要占用更多空间。如果预留很大、实际只用一小部分,昂贵芯片里就会堆积无人愿意买单的空白。把内存想成图书馆书架,就会明白问题不一定是书太少,也可能是抽屉分配方式不对。

2023 年,研究人员发表了 vLLM 系统及其分页注意力算法 (PagedAttention)。它借鉴操作系统的虚拟内存分页 (Virtual Memory Paging),把注意力缓存切分为可以灵活管理、共享的区块,减少连续预留造成的浪费。在同行评议论文的实验中,与特定基线系统相比,该方法在近似延迟水平下把推理吞吐量提高到约两至四倍 [S05](#)。这不意味着任何数据中心装上软件都能复制同样的增幅,却证明一个算法决策可以重新分配现有硬件的价值。

更早的 Orca 系统在 OSDI 2022 上展示了按单次生成迭代调度 (Iteration-Level Scheduling) 的方法,可以减少短请求被更长批次阻塞的等待 [S06](#)。连续批处理 (Continuous Batching)、前缀缓存 (Prefix Caching) 和两阶段计算分离,不是装饰性的附加功能;它们决定有多少计算能在可接受时间内变成服务。量化 (Quantization) 降低数值表示精度,从而改变内存和计算开销,但必须进行质量测试,因为节省的资源可能由用户在回答准确性上付出代价。推测解码 (Speculative Decoding) 先用更快的函数提出候选令牌,再由另一个模型校验,因此不同负载下的加速比例不会恒定。

所以,理论峰值浮点运算次数 (Floating-Point Operations per Second; FLOPS) 不能等同于客户每小时愿意购买的正确答案数。两者之间隔着队列、模型、调度器、能源、响应时限合同,以及长短不一的问题。效率不是芯片的固定属性,而是一整套系统执行某项具体工作时呈现出的行为。

4. 悖论: 软件既能挽救投资, 也能同时压低它的售价

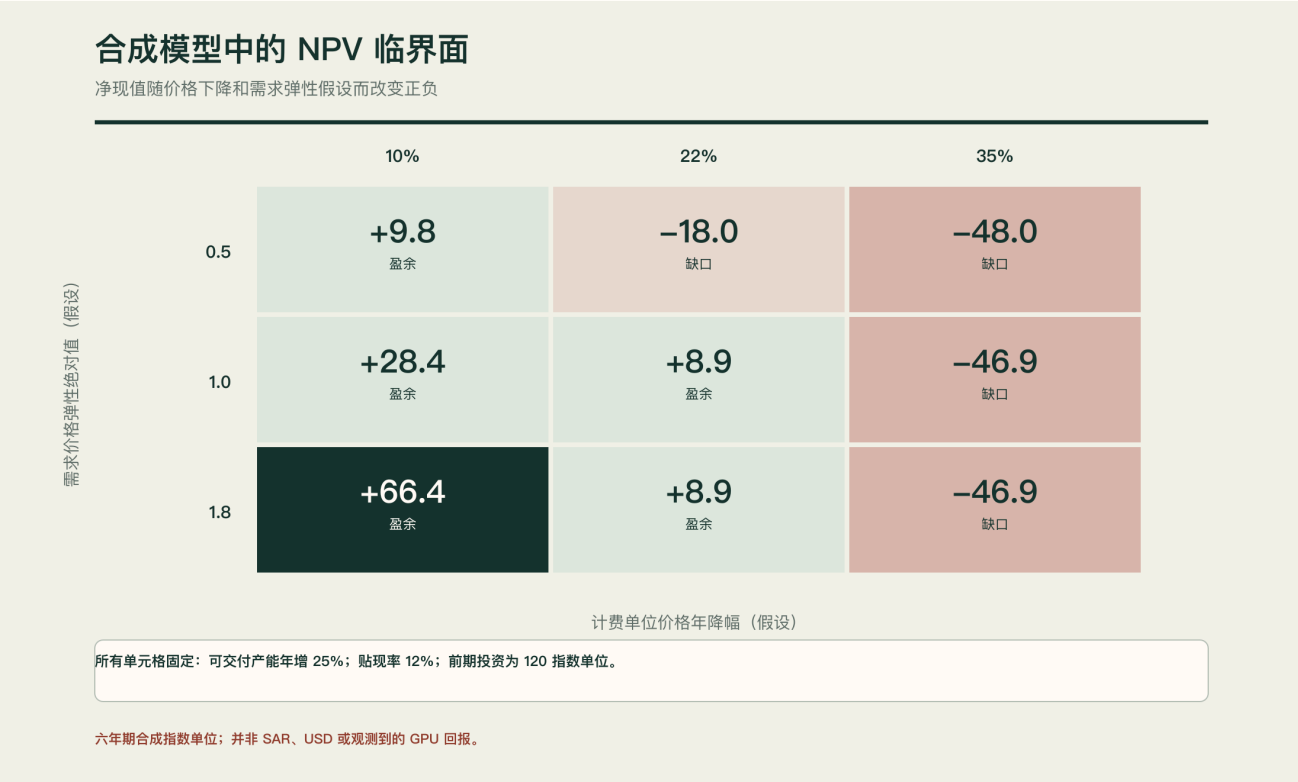
设想一个纯属假设的例子,不要把数字归于沙特或任何公司。一台服务器原本每天完成五万个工作单位;内存管理改进后,不换设备就能完成八万个。这是漂亮的生产率增长。但如果所有竞争者都能部署同样的改进,算力不再稀缺,市场价格就会被重新谈判。用户同时受益于速度和竞争;服务器所有者却不一定能同时拿走两种收益。

判断结果需要看经济学所说的需求价格弹性 (Price Elasticity of Demand): 价格下降时,需求量会增加多少? 用一个简化形式表示,若需求与价格的负 ϵ 次方成正比,那么在容量与成本限制之前,收入就与价格的 $1-\epsilon$ 次方成正比。 ϵ 大于 1 时,降价后总收入可能上升; ϵ 小于 1 时,即使消费增加,收入也可能下降。容量一旦满载,需求再旺盛也没有意义,除非能够真正交付付费单位。

这是威廉·斯坦利·杰文斯 (William Stanley Jevons) 1865 年在《煤炭问题》 (The Coal Question) 中提出问题的精确变体:效率提高,会不会反而增加资源总消耗? 杰文斯并未预言人工智能令牌。更不能把这段历史观察写成“任何改进都会必然让需求翻倍”的口号。反弹效应 (Rebound Effect) 只是一种可能,

它由弹性、可用容量、替代应用的价格和已完成工作的价值共同决定。如果这段历史只留下一条规则，那应该是：不要根据工程师有多喜欢一颗处理器来推断需求曲线 S08。

市场因此需要同时回答两道方向相反的问题。软件会不会推迟采购下一代设备，从而提高现有设备的回报？它又会不会让服务充裕到足以压低所有人的售价？两道题完全可能同时答“是”。判断谁获利，不能只数服务器吐出多少令牌，要看卖出这些令牌后还剩下多少现金利润。



按需求弹性和降价幅度绘制的回报图；这是说明性模拟，不是公司数据

图 1. 以假设单位计量的净现值（Net Present Value; NPV），并非美元或里亚尔。图中组合三种降价幅度与三种需求弹性，并假设可售容量每年提高 25%。每个单元格都来自随附模型的实际运行，不是市场预测。

5. 当利用率指标损害了我们对财务的理解

投资者有时像酒店老板看入住率那样看 GPU 利用率（GPU Utilization）。这个比喻有用，直到我们记起：酒店客房不会因为一次软件更新，在夜间从套房变成走廊里的一把椅子。处理器处于运行状态，也不代表全部算力都能计费。资源可能花在数据准备、失败重试、网络与内存之间的等待、客户不购买的内部工作，或质量不足、无法交付的执行结果上。

一颗利用率 95% 的处理器，收入可能低于另一颗利用率 65% 的处理器，原因可以是价格、请求价值、不可计费浪费或服务等级不同。反过来，成功的软件改进也可能令利用率下降，因为真实容量增长快过需求；此时下降并不代表运营失败。问题出在分母，不是关于勤奋与懒惰的道德故事。

在随附计算模型中，假设初始年服务能力为 50 个单位，单位价格为 1；资本成本为 120 个会计单位，分析窗口为六年，年折现率 12%，并明确设定固定与可变运营成本。这些数值不来自任何具体公司的账目。它们的作用，是让模型以修辞无法做到的方式反驳直觉。

当假设价格每年下降 22%、有效容量提高 25%、需求弹性为 0.5 时，净现值为约 -18 个单位。弹性升至 1.0，净现值变为约 +8.9 个单位。把弹性继续提高到 1.8，在这组假设下也不会进一步提高结果，仍约为 +8.9，因为两种需求都撞上可交付容量上限 [S12](#)。市场外面就算排满客户，也不会为数据中心无法提供的单位付钱。

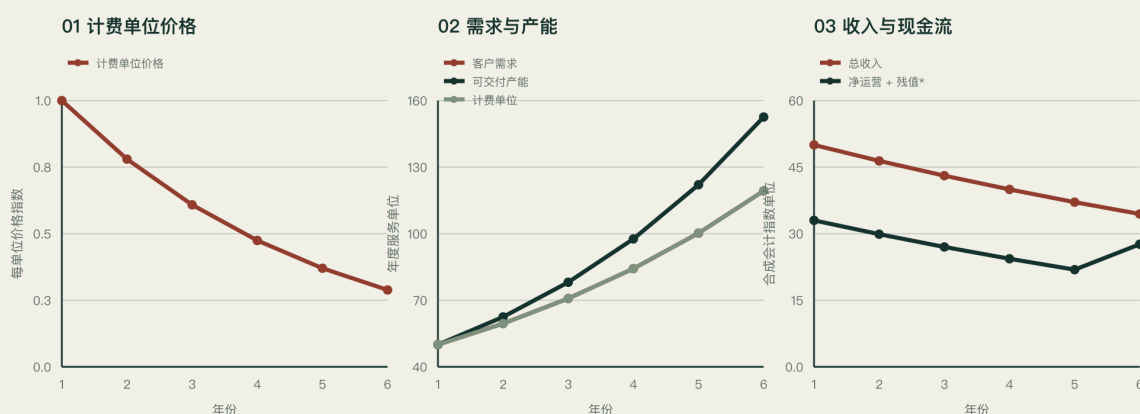
我们还运行了一个由 27 个情景组成的网格，组合三种降价率、三种弹性和三种产能提升速度。结果有 12 个正净现值情景，15 个负净现值情景。这个计数不代表盈利概率；所有单元格都是为测试而预先设定的合成情景。更重要的结论是，两句看似有力的话都不能成为普遍规律：“廉价智能会吃掉芯片回报”，以及“需求增长会挽救所有投资”。两者都需要明确、可测的条件 [S12](#)。

利用率之外，还有一个更漂亮的错觉。基础模拟中，现金流在第五年名义收回全部资本成本，净现值却仍然约为 -5.6 个单位。这不是计算矛盾。资本有价格，五年后回来的钱，不等于今天流出去的钱。一个项目可以庆祝“回本”，同时在既定折现标准下仍不如替代方案有吸引力。一张长期合同的照片，回答不了等待成本。

还必须把模型中的乐观假设摆到台面上：我们让单位可变成本随着容量改善以相同比率下降。在独立敏感性测试中，若可变成本保持不变，而基础情景其他条件不变，净现值就会从 -5.6 降至 -31.2 个单位。因此，一份讲“生产率提高”的精美演示，不能替代对每项服务实际耗电与运营开支的测量。完成更多工作更快，不代表每一项工作都更便宜 [S12](#)。

价格与产能的分离

基准情景： $\delta=22\%$ ； $\epsilon=0.7$ ；可交付产能每年增长 25%



* 第 6 年净现金流包含假设的 8 单位残值；第 0 年的 120 单位初始资本开支未显示。

假设模型输出：没有观测到的产能、价格或项目现金流。

六年间价格、规模与收入的变化；假设示例

图 2. 同一合成财务实验中的三组序列。容量的物理变化与收入的商业变化逐渐分离；需求表示市场希望购买的数量，而非实际交付量。完整坐标范围与参数见研究附件。

6. 沙特：从一排排机架，走向真实客户经济

2026 年 8 月 31 日，AMD、思科（Cisco）与 HUMAIN 宣布，一套部署在沙特、采用 AMD Instinct MI355X 处理器、Cisco Silicon One 网络和 ROCm 软件的计算基础设施已经投入运行并为客户提供服务。三方还宣布，计划从 2027 年开始部署最高 250 兆瓦容量，并以 2030 年达到一吉瓦为扩张目标。这是三种不同状态，不能写成一句话：按项目方声明已经运行的能力、计划中的未来扩张，以及依赖执行、需求与融资的远期目标 [S09](#)。

这条消息让沙特的的问题更有层次。AMD 处理器已经进入可能的付费服务，不再只有采购英伟达设备这一种叙事。由此可以比较不同硬件家族的总体拥有成本（Total Cost of Ownership; TCO）、软件兼容性（Software Compatibility）、模型可迁移性（Model Portability）和真实应用表现。但消息并不能证明哪一个平台更便宜或更赚钱，也没有披露客户合同、利用率或净利润率。这些都需要公告之外的数据。

再往前一年，HUMAIN 宣布推出基于 ALLaM 34B 模型的阿拉伯语聊天应用，功能包括阿拉伯语语音输入、搜索，以及阿拉伯语和英语使用。公告没有告诉我们服务了多少令牌，也没有披露每次会话的收入；它却让计算中心从一座能源模型，变成公民手机上或机构工作中可能发生的使用。真正关键的经济转换就在这里：计算机并不会因为“会思考”就自动形成市场；只有个人或机构发现某项工作值得交给它，并愿意为此花钱，市场才会出现 [S10](#)。

如果沙特希望从购买基础设施进一步走向推理产业（Inference Industry）的竞争位置，就不该在“最大的数据中心”和“最便宜的令牌”之间二选一，仿佛选择到此结束。更关键的是，找出能够以可预期质量交付、并且成本可以随时间下降的高价值工作。它们可能是针对阿拉伯语优化的流程、经过准确性测试的专用模型、高隐私处理托管服务，或保证响应时间的模型开发服务。这些只是算力经济中的类别示例，不是在宣称它们已经存在或必然盈利。

7. 一场测试，可能让五颗相似的处理器的变得截然不同

技术公告很擅长把巨大差异压进一个词：“性能”。然而，同时回答五个短问题的推理引擎，与承载一千段共享上下文的长对话时，并不是同一个引擎。首令牌时间（Time to First Token; TTFT）不同于令牌间延迟（Inter-Token Latency; ITL），两者又不同于每秒处理令牌数（Throughput）。一台设备可能赢得令牌速度比赛，却因首个回答来得太慢而输掉强调即时响应的应用；另一台设备可能单请求速度排名靠后，却在高并发中胜出。

2026 年 9 月 16 日，MLCommons 发布 MLPerf Inference v6.1 更新，其中加入更接近多步骤任务的新测试，也纳入按基准规则审核的新系统结果。这是一项有价值的知识增量，但不是任何厂商的盈利证书。基准测试衡量的是规定条件下的运行表现，不衡量算力运营商向客户收取的实际价格、融资、税费，也不衡量设备闲置成本 [S11](#)。

不妨设想两台同代设备的真实竞争，暂时不知道谁更好：

- 第一台在标准负载下每秒输出四千个令牌，但每位客户的月度运营成本很高，因为必须预留大量容量才能保证首令牌速度。
- 第二台在同一负载下每秒只能输出三千五百个令牌，却能共享上下文内存，高效处理长短不一的批次，并复用用户的重复请求，因此可以在客户付费的时限协议内服务更多人。

以上是说明比较方法的虚构示例数字，不是对两台公开设备的测量。只有固定模型、输入和输出长度、回答质量、真实请求比例、延迟上限、电力冷却与运营成本，以及可计费工作比例后，才能宣布赢家。届时，“每项正确且被接受任务的成本”（Cost per Accepted Task）或“符合服务等级协议任务的利润”（Margin per SLA-Compliant Task），会比孤立比较硬件计算次数更接近经济现实。

问题就此离开处理器机箱，来到投资经理的办公桌：究竟哪一组工作负载适合这台设备？客户需要长时间连续生成时，内存与传输带宽会成为约束；大部分使用是在共享企业文档中检索时，上下文复用可能更重要；客户需要训练新模型时，价值分配与换代敏感度又会改变。把所有账单都塞进一台名叫“小时价格”的计算器，再对结果不同感到意外，并不聪明。

8. 电力公司无法用一项技术妙招修好销售合同

算力世界里有三种很难同时放进一张图的速度：新一代硬件制造的速度、电力容量安装的速度，以及发现愿意为结果付费的需求的速度。第一种跟随产品和研究周期，第二种跟随建设、电网和供应链，第三种则跟随机构试验——它们可能热情开场，后来却发现并不需要持续推理。三种速度之间的缝隙，产生了实物期权（Real Option）的价值：可以推迟设备交付，调整工作负载类型，或者先租用阶段性容量，而不是立刻把所有未来雄心变成今天的账单。

再看一类不会出现在数据中心宣传图里的合同。CoreWeave 表示，其许多多年期合同采用“照付不议”（Take-or-Pay）安排；截至 2025 年末，公司已承诺合同的加权平均期限约为五年。按照公司对这些合同的描述，即使客户没有完全使用容量，也须依约为已承诺容量付款。这种经济机制可以把一部分利用率下降风险从运营商转移出去。因此，不能把人工智能接口降价对其收入的影响，推演成所有服务都按公开即时价格随买随卖 [S01](#)。

合同并非魔法护盾，也不会让时间消失。实际效果取决于交付能力、客户权利、定价与续约条款、各方偿付能力以及债务支付日期。大客户即使不再需要所预订的算力，也可能继续付款至合同结束，再在续约时强硬谈判。算力中心若遇到更高价的盈利机会，也未必能把已经承诺出去的容量重新分配。环境逆转时，连续性带来的利润保护，可能变成机会成本。不过，我们并不知道沙特合同的细节，不能把一家美国公司的公式直接套上去。

判断处理器回报，关键在盈余和缺口出现的时点，而不只是现金流总额。一笔贷款在需求成熟前到期，项目即使后来净现值为正，也可能先承受现金压力。固定价格的长期合同能保护现金流免受降价冲击，却也可能阻止运营商充分分享价格上涨。强有力的软件改进甚至可能同时提高单位利润、降低单位收

入。不能把这些权衡压成一份“治理风险”清单；每一项都是可计算的财务关系：还款表、需求曲线、用量上限、技术换代时点和电力容量窗口。

9. 一个战略设想：不要只拥有处理器，要拥有它的经济学习曲线

到 2030 年，算力运营商最重要的资产，也许是参观数据中心时镜头拍不到的东西：一份精确历史记录，把工作负载类型与实际资源消耗、客户付款内容和客户接受的质量连接起来。它不是美化后的调用次数统计，而是每个月积累下来的知识：时间在哪里流失，利润又在哪儿形成。

我把这项研究设想称为算力商业寿命图谱（Compute Commercial Life Map）。对每一种工作负载，分别记录五个时间：让模型适配硬件所需的时间、首个输出前的等待时间、得到可接受输出的时间、价格仍能在竞争中站得住脚的时间，以及该负载分摊投资的现金回收时间。五种时钟并不相同，也不能全部压缩成一个全国平均值。芯片的物理寿命可能是六年，价格优势可能只有八个月；在较小模型上的有效寿命可能长得多，而占用它的合同期限又与前三者没有直接关系。

设备销售者通常会拿最好的基准测试说服你。经验丰富的投资经理更想看到经济优势衰减曲线（Economic Advantage Decay Curve）：一种服务在需要降价或更换工作类型以前，能比替代方案多保持几个月的价值？这应按设备批次测量，而不是按处理器世代口号测量。这样才能比较三项不适合作为竞赛标题的行动：优化软件、重新定价服务，或者改变适合该算力的客户群。一年中最聪明的算力决定，也许是不再加购一批设备，只是它不适合作为耀眼的头图。

一个反对悲观结论的赌注也值得认真对待：推理成本下降，可能催生过去根本无法出售的应用；即使价格降低，市场总量和盈利仍会上升。新的复杂工作也可能吸收全部效率增量，而不是触发价格战。沙特还可能受益于能源条件、投资能力、区域市场和阿拉伯语应用。然而，这些是条件和待检验的路径，不是可以从项目宣称规模直接推导的结果。没有付费需求、单位利润和已执行合同的数据，两种故事都仍然成立。

10. 下一则公告发布前，值得先读懂的五件事

只要有获准使用、经过汇总的运营和商业数据，研究人员或投资负责人明天就可以开始，不必等待一个不存在的万能指标：

测量对象	它揭示的问题	只看最容易取得的数字会犯什么错
可接受任务收入 (Revenue per Accepted Task)	客户付款是否高于成功完成任务的全部成本?	不论任务差异, 只看每百万令牌价格。
服务等级约束下的有效吞吐量 (SLA-Constrained Throughput)	在约定质量与延迟下能完成多少请求?	使用处理器实验室峰值。
合同锁定收入占比 (Contracted Revenue Share)	有多少现金流不依赖即时价格, 具体条件是什么?	把合同公告等同于已收现金或保证利润。
经价格调整的回收期 (Price-Adjusted Payback)	项目能否在商业优势尚存时收回资本?	只看会计折旧寿命。
工作负载利润分布 (Workload Margin Distribution)	平均值是否掩盖了亏损负载与补贴它们的盈利负载?	展示漂亮的总利用率, 抹平应用差异。

可以设想一场在沙特市场公开进行的测试: 两家服务商使用匿名、相近的工作负载, 在统一时限内竞争提供质量已知的阿拉伯语服务。测量双方每项可接受任务的完整成本, 然后在需求规模、对话长度分布和上下文复用比例变化时重复试验。再加入经济测试: 服务价格降低 20% 后, 新获得客户、完成请求和利润分别怎样变化? 无需泄露关键设施细节或客户秘密, 只需公开汇总后的比较结果和指标定义。结果不应预设: 最新设备、最好的软件、最合适的工作, 或最均衡的合同, 都可能胜出。

本文提出的假设是: “一个国家在人工智能算力上的优势, 不来自拥有规模最大的抽象容量, 而来自在技术改进贬值以前, 把它更快转化为有用且付费的工作。”这句话应当接受检验。如果控制工作负载、成本和合同条款后仍找不到关系, 就应放弃或缩小它。我不想用一个漂亮比喻凭空建立市场。

11. 如果今天把问题问对, 2030 年会告诉我们什么?

第一种情景中, 复杂推理需求增长快过价格下降; 每一代新硬件发布, 都打开旧设备无法支撑的新应用, 沙特算力由更大的区域市场获益。第二种情景中, 软件和新计算机制造出巨大供给, 但大部分价值转移到应用开发者与用户手中; 算力中心运营商转向激烈的价格和成本竞争。第三种情景中, 两类市场并存: 少量高价工作使用顶级容量, 通用算力则成为低利润商品。它们是没有概率权重的条件路径, 不是三项经过计算的预测。

2026 至 2030 年间应该观察什么? 不是只看公开标价, 而是看实际成交价格、已承诺合同的真实占比、输入输出与缓存结构、统一时限下的服务表现、同一负载上不同硬件世代的任务成本差异, 以及来自本地市场之外客户的收入占比。如果令牌量增长快过收入, 不必立刻宣布灾难; 先检查利润率, 再看资产回收, 最后看获客成本。收入增长也不等于胜利; 还要看资本、电力冷却和网络吞掉了多少。

过去的工厂经理听机器的节奏, 知道沉默可能意味着损失。现代智能工厂的经理需要另一种勇气: 看着处理器产出更多, 然后礼貌而严厉地追问, 究竟是谁为这些产出付了钱, 多少工作没有计费, 又有多少客户其实可以用更小的设备完成同样计算。

现代机器可以赢下每一场速度测试，却输掉下一次定价会议。旧一代也可能在性能榜上落后，却因为它执行着熟悉客户、熟悉需求的工作而继续赚钱。两者并不矛盾。真正的矛盾，来自我们把科学进步和账单进步混为一谈。

读者最后应当记住的问题，不是：沙特会有多少颗处理器？而是：沙特将比别人更擅长计算什么？谁又会在下一代处理器到来以后，仍然需要这项计算并愿意为它付费？

来源与延伸阅读

1. [CoreWeave, 2025 Form 10-K, SEC: 使用寿命与合同信息, 2026 年发布。](#)
2. [CoreWeave, Form 10-Q, period ending 30 June 2026, SEC: Note 2 与 RPO。](#)
3. [NVIDIA, Rubin platform announcement, 5 January 2026。](#)
4. [OpenAI, Introducing GPT-6 Sol and Luna, September 2026: 定价与缓存部分。](#)
5. [Kwon et al., Efficient Memory Management for Large Language Model Serving with PagedAttention, SOSP 2023, DOI 10.1145/3600006.3613165。](#)
6. [Yu et al., Orca: A Distributed Serving System for Transformer-Based Generative Models, OSDI 2022。](#)
7. [NVIDIA, TensorRT-LLM, Disaggregated Serving technical documentation。](#)
8. [William Stanley Jevons, The Coal Question, 1865, digitized historical edition。](#)
9. [AMD, Cisco and HUMAIN, joint production and expansion announcement, 31 August 2026。](#)
10. [Saudi Press Agency, HUMAIN Chat powered by ALLaM 34B, 25 August 2025。](#)
11. [MLCommons, MLPerf Inference v6.1 Results, 16 September 2026。](#)
12. [可复现的原创合成模型、假设与 27 个情景输出；另见结果表与假设。](#)

编辑说明：公开财务数字均为注明出处的来源事实；模拟结果不是沙特公司或服务的实测数据。文中所述机构没有参与本文的撰写或审阅。编辑日期：2026 年 10 月 24 日。中文稿为待内容所有者审阅的编辑草稿；不声称经过独立中文母语人工审核。