下载日志已记录，仅供内部参考

证券研究报告|策略深度报告
产业经济
2026年8月15日

华福证券 HUAFU SECURITIES CO.,LTD.

中国致力于打造全球AI公共产品底座，或将引领全球算力走向冗余时代

证券分析师：
李 浩　执业证书编号：S0210524050003
李刘魁 执业证书编号：S0210524050006

下载日志已记录，仅供内部参考

核心观点

- **核心要点：当AI算力被定义为全球公共品，中国的战略资本就不再是"补贴"，而是基础设施投资的天然属性；美国的金融资本回报逻辑，在这一框架下面临根本性重估。**
- 美国Token财政本质：美国版"土地财政"——算力/Token垄断性涨价为锚、CSP寡头替代政府财政、企业部门（to B）替代居民部门；债务驱动的通胀生成机制，而非科技革命叙事
- 中国三大优势：①模型效率——中国开源模型以更少的激活参数产出同质量token，单位算力消耗更低；②资本开支——同等规模数据中心建设投入远低于美国；③融资成本——政策性资金支持下的加权资本成本明显更低。
- 更关键的是信息不对称：中国模型开源、参数与效率可验证，美国模型闭源、效率无法核实。在Token价格竞争中，可量化的成本优势比不透明的技术优势更具说服力。
- 未来优势将进一步放大：中国当前唯一短板是国产芯片能效偏低，拖累单位电力的token产出。一旦硬件约束解除，无论是国产芯片追赶还是先进制程可得，产出效率或将成倍提升，盈亏平衡价格或进一步下探，对美国最优方案仍保持数倍优势。即中国的成本优势并不依赖硬件代差缩小，自主可控只是把"成本领先"升级为"成本+性能双领先"。
- 从资本回报品到普惠型基础设施：AI算力首先应服务于人，而不应仅以高毛利和高资本回报为唯一目标。
- 从单边垄断到国际公共产品供给：人工智能不应是某个国家的独奏，而应是全球合作的交响。
- 中国AI供给的扩张不只改变国内竞争，也将通过更低成本的模型、算力和应用方案影响全球token价格中枢。
- 风险提示：美国AI应用与生产率兑现超预期、Token价格提前见顶回落、AI支出向传统工业品转移不及预期、成本测算关键假设存在不确定性、国产算力进程与中美产业政策超预期变化

下载日志已记录，仅供内部参考

目录

- AI即宏观：Token财政与美国经济新范式
- AI算力=制造：中美资本配置逻辑与竞争底层差异
- 中美1GW数据中心成本对比测算
- 普惠供给与全球竞争重估
- 风险提示

下载日志已记录，仅供内部参考

研究视角：AI即宏观

- 当AI融资规模跨越万亿美元量级，其对信用结构、通胀传导与全球流动性分配的影响已无法被传统宏观框架忽视。美国AI，就是当下的美国宏观。
- 我们的研究路径是：沿着AI融资所派生的债务和资本开支，追溯资金来源、分配去向、通胀载体和全球影响，从而构建一套可以解释乃至预测美国未来经济运行规律的宏观框架。
- 我们将当前美国AI驱动的经济模式定名为“Token财政”。其核心逻辑如下：
  - 以算力/Token的垄断性涨价为锚，维系债务循环不断扩张；
  - 以CSP寡头联盟替代政府财政，承担基建投资功能；
  - 以企业部门（to B需求端）替代居民部门；
  - 通胀路径从算力涌现，向传统工业蔓延，带动更广泛的库存周期。

下载日志已记录，仅供内部参考

什么是Token财政?

- **美国联邦政府受债务上限约束，传统财政扩张能力有限，无法大规模主导基建投资。**
- **Token财政是美国版“土地财政”**：以算力垄断性涨价为锚，企业部门替代政府财政执行信用扩张和基建投资。
- **土地财政的三要素**：供给约束+需求加杠杆+信用扩张，Token财政逐一对应
- **关键区别**：执行主体从政府变为CSP（云服务提供商，后统一简称CSP），现阶段买单主体从居民变为企业（to B）

美国版“土地财政”

维度	中国土地财政	美国Token财政
通胀锚	土地价格	Token/算力价格
执行主体	地方政府	大型CSP
买单主体	居民（房贷）	企业（to B订阅）
基建功能	政府财政直投	CSP协议性出资
信用载体	按揭/城投债	私募信贷/IG债
脆断点	房价需稳	Token不能通缩

资料来源：华福证券研究所绘制

下载日志已记录，仅供内部参考

Token财政循环链条与脆断点

- **循环自洽，但前提是Token持续通胀；一旦价格拐点出现，链条同步断裂.**
- 正向循环：涨价→融资→基建→企业盈利→to B买单→涨价维持
- 脆断点识别：类比2007年美国次贷危机“房价需稳”，当前 “Token不能通缩”是等价命题

美国版“土地财政”

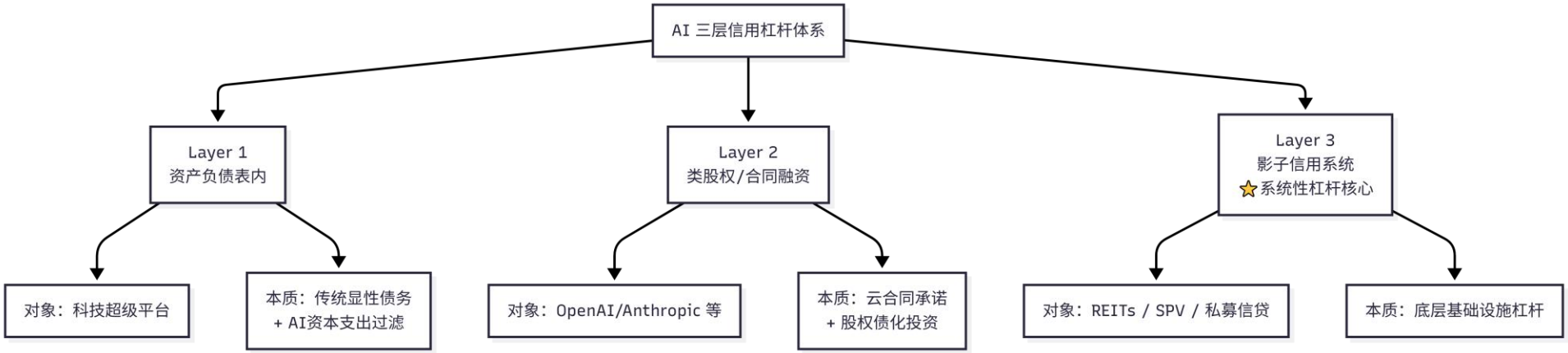
Token 财政循环链条
Token 涨价 → CSP 超额利润 → 偿债能力增强 → 融资成本下降 → 资本开支扩张
资本开支扩张 → 电力/基建需求上升 → 工业通胀 → 企业部门收入扩张 → to B 买单能力增强
to B 需求旺盛 → Token 用量增加 → Token 价格维持 → 循环继续
⚠ 脆断点：一旦 Token 价格出现下行拐点，融资成本骤升，债务偿还压力急剧扩大，循环断裂。

资料来源：华福证券研究所绘制

下载日志已记录，仅供内部参考

AI债务的规模

- **伴随AI资本开支扩张，AI相关的融资模式已脱离了传统的单一公司发债模式。**
- **我们可以将AI融资刻画为：AI融资 = 显性债务 + 合同化债务+ 影子信用扩张**



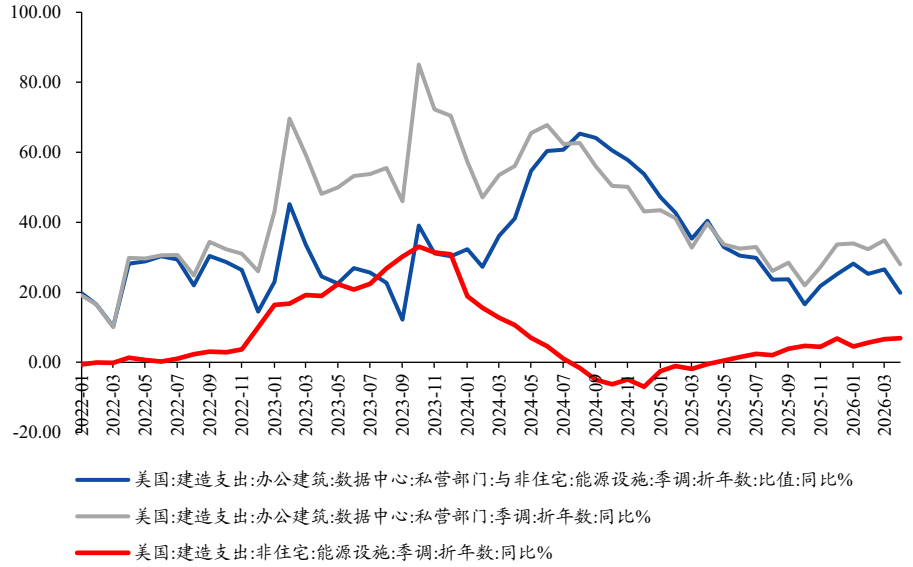
资料来源：华福证券研究所绘制

下载日志已记录，仅供内部参考

从算力到基建：历史性的支出权重转移

- AI资本开支权重正从“纯算力”向“算力+传统基建”转移
- 这不是算力投资减少，而是总盘子扩大后基建权重崛起

2026年以来，美国数据中心建造增速边际走弱，能源设施增速走强



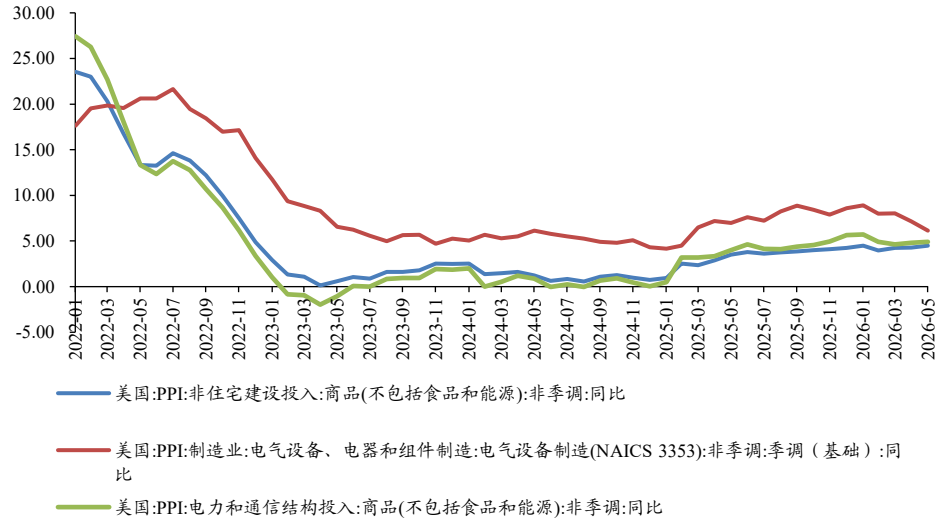
资料来源：wind，华福证券研究所

下载日志已记录，仅供内部参考

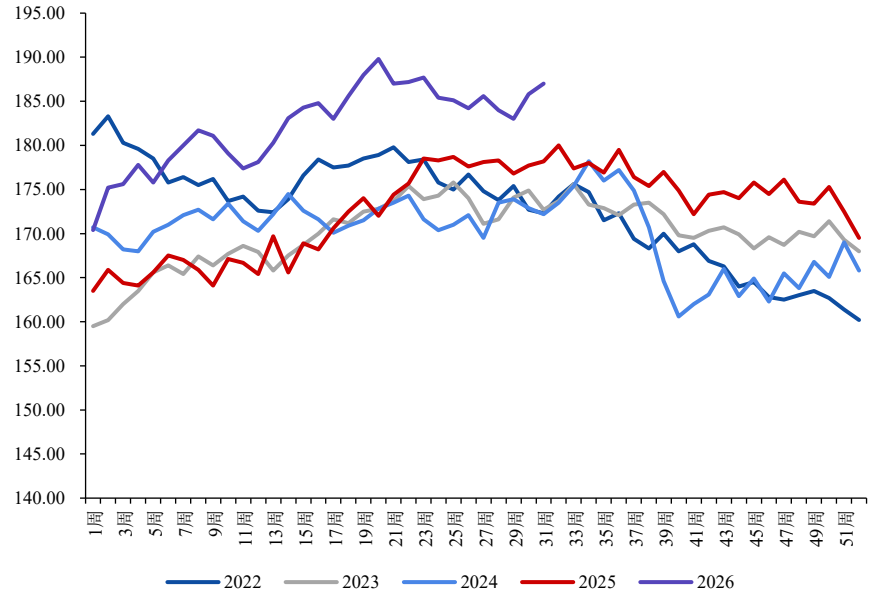
基建投资细分：电力基建影响较大

- 基建支出构成中电力基建影响已经显现，美国PPI建筑子项已开始验证。
- 能源/电力→ CSP自建电源，与电力公司签PPA（企业购电协议）
- 厂房建设 → 钢铁需求激增

美国PPI非住宅建设和电力投入同比增速上行（%）



美国粗钢产量（万短吨）



资料来源：wind，华福证券研究所

下载日志已记录，仅供内部参考

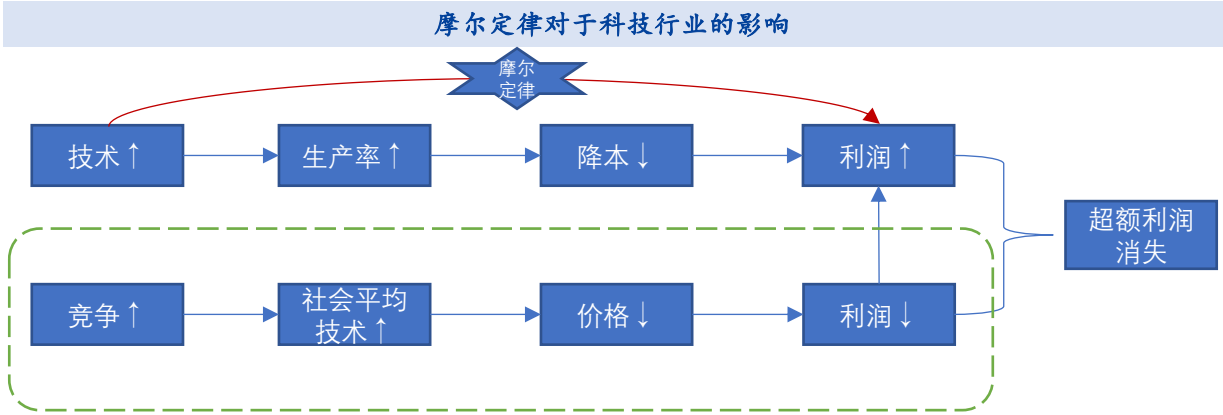
目录

- AI即宏观：Token财政与美国经济新范式
- **AI算力=制造：中美资本配置逻辑与竞争底层差异**
- 中美1GW数据中心成本对比测算
- 普惠供给与全球竞争重估
- 风险提示

下载日志已记录，仅供内部参考

AI算力是制造业

- **AI算力=IT硬件+传统基建**
- **IT硬件：摩尔定律失效后，IT硬件向制造业属性回归**
- **传统基建：离不开制造业**



资料来源：华福证券研究所绘制

下载日志已记录，仅供内部参考

制造业竞争的核心是要素成本

- **AI算力的核心投入要素：土地、电力、建筑、硬件设备，全部是制造业要素。Token的生产成本由以下四项决定：**
- **IT硬件折旧（GPU等）**
- **基础设施折旧（电力/冷却/建筑）**
- **电力运营成本（电费）**
- **融资成本（WACC）**

- **回顾制造业的全球迁移史：**纺织业从英国到日本到中国，钢铁从美国到韩国到中国，半导体封装从美国到中国台湾到东南亚，制造业总是向要素成本更低的地方迁移。这一规律同样适用于AI算力。

下载日志已记录，仅供内部参考

资本配置方式决定竞争格局，历史规律正在重演

- **在制造业竞争中，战略资本驱动有系统性优势——它可以在金融资本主义认为"不经济"的价格上持续投资，直到规模效应出现。AI算力正在经历这个过程。**

- **美国模式：金融资本驱动**
- 资本配置由市场价格信号（利率、股价、ROIC）驱动，投资决策以净现值最大化为准则。资本的时间偏好由市场利率决定，一旦预期回报低于资本成本，投资停止。季报周期、信用评级、股东回报压力构成核心约束。
- 这套逻辑在轻资产、高利润率的软件和平台业务中运转良好，但在资本密集型的制造业中，往往产生系统性的投资不足——因为制造业的ROIC在规模效应出现之前，通常低于金融资本的回报要求。

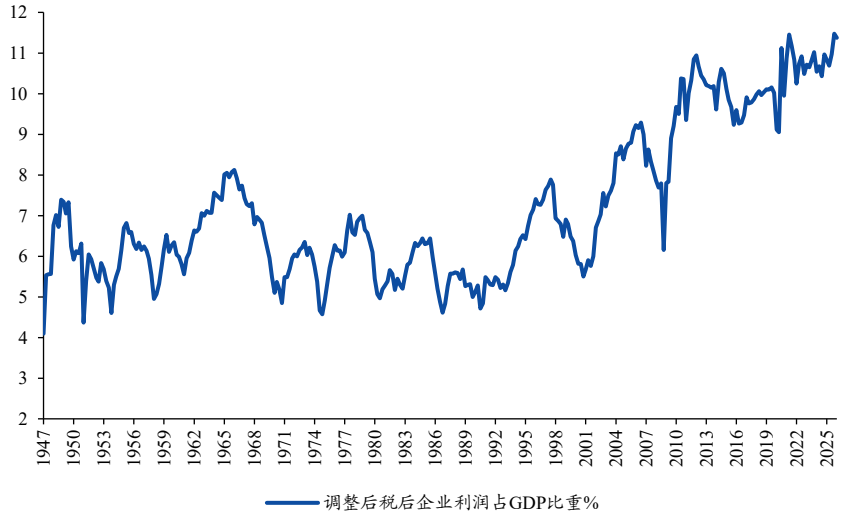
- **中国模式：战略资本驱动**
- 资本配置以国家战略目标（产业升级、技术自主、供应链安全）为导向，政策性金融工具提供低成本、长周期的资本支持，融资成本较低。
- 投资决策的时间维度与五年规划周期对齐，容忍更长的回报周期，允许在规模效应出现之前持续投入。

下载日志已记录，仅供内部参考

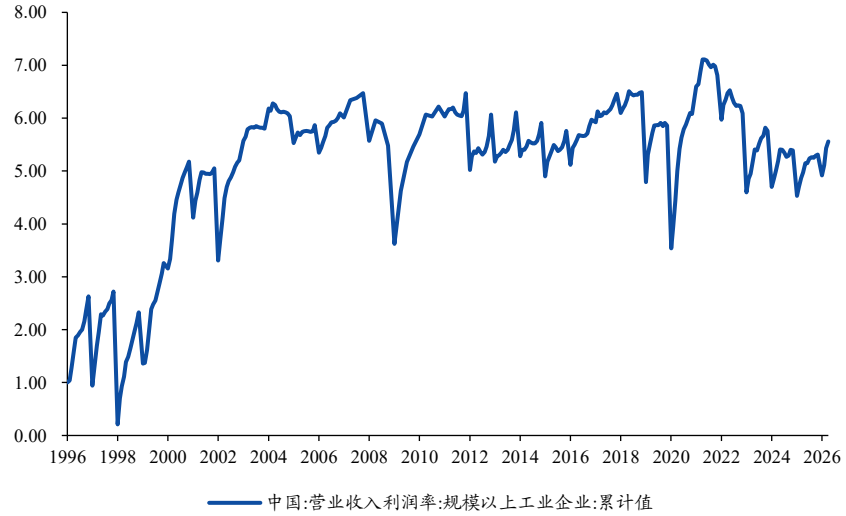
金融资本主义驱动高利润和高支付力

- **金融资本逐利的微观行为助推美国宏观利润率的提升，高利润率对应了高支付能力。**

美国企业利润率（%）



中国企业利润率（%）

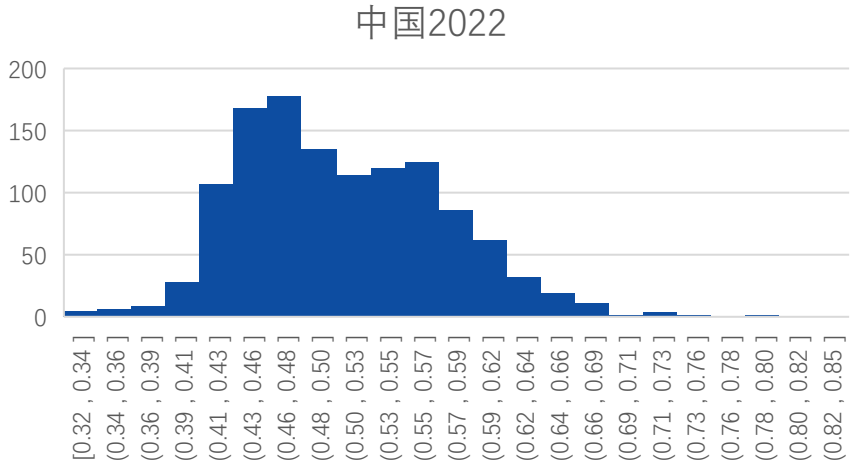


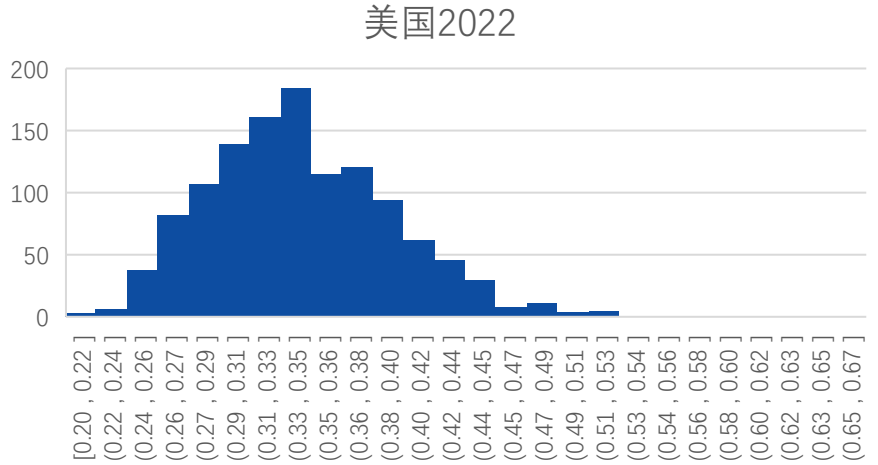
资料来源：FRED、wind，华福证券研究所

下载日志已记录，仅供内部参考

中国的战略在于构建全球领先的制造业完整度

- **“舍弃”资本利得后，中国获得了制造业的完整。**
- 以下图表刻画的是中美制造业的完整度：横轴是中美生产并出口商品与本国制造业相关系数区间，纵轴是商品种类的数量，高相关性区间的商品种类越多，就可以刻画出一个国家的制造业完整度较高。



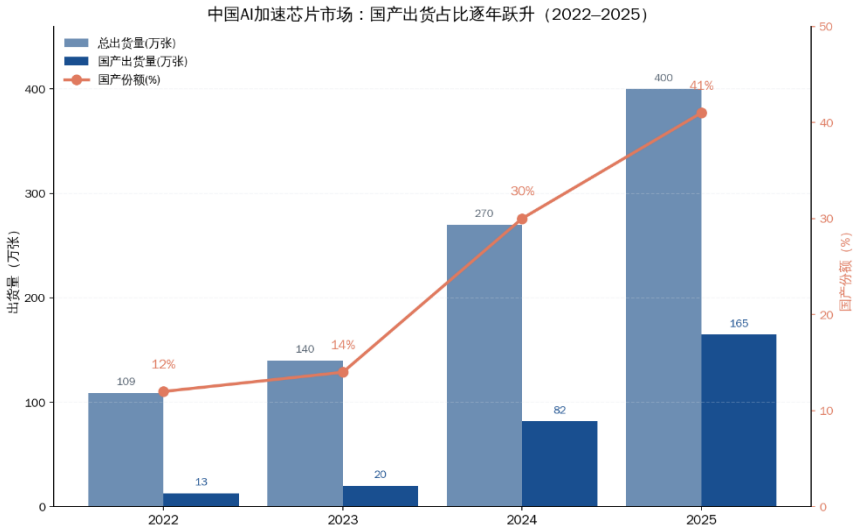


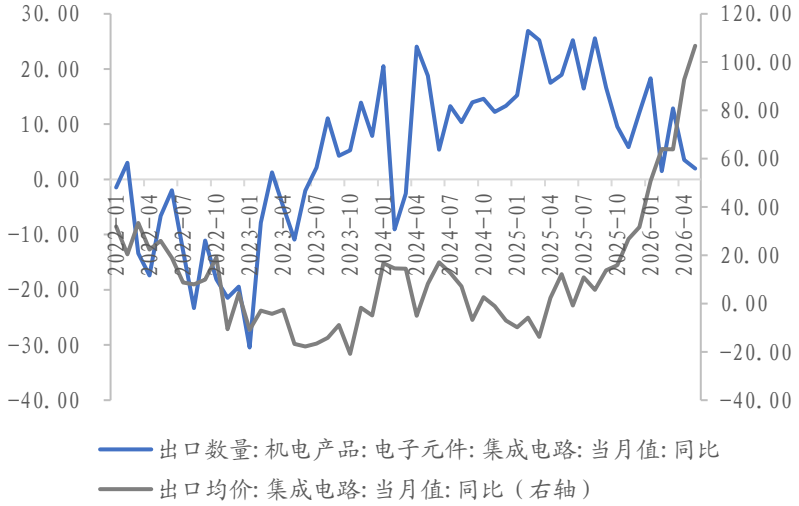
资料来源：OEC，华福证券研究所

下载日志已记录，仅供内部参考

自主可控的进程

- **自主可控不是口号，是可以量化的进程**
- **昇腾910C（2025）：H100性能的约60%**
- **自主可控的进程是非线性的，但方向是确定的。**





资料来源：智通财经网、中研网、新浪财经、IDC、reuters，华福证券研究所

下载日志已记录，仅供内部参考

中国如何平滑劣势

- **ToB支付能力弱**
- **解法：开源模型大幅降低使用门槛**
- DeepSeek、Qwen、GLM全部开源。企业可以在自己的服务器上部署，在按 Token 付费的 API 模式下，企业业务量越大、调用频次越高，AI 成本就呈现线性甚至指数级上升（OPEX 纯变动成本）。而本地部署后，模型的软件授权本身免费，成本转化为硬件设备与算力的投入（CapEx 资本支出 + 固定折旧），大幅降低了中小企业的门槛。
- **创新激励不足**
- **解法A：允许民营科技企业在特定领域充当"创新引擎"**
- 中国并不是纯粹的国家资本主义：华为、阿里、腾讯、字节跳动、DeepSeek都是民营企业，在技术创新上受市场竞争激励。国家资本提供的是基础设施和资金，创新本身由民营企业完成。
- **解法B：竞争性政策，不允许单一企业垄断**
- 中国AI芯片有华为、寒武纪、百度昆仑、阿里平头哥、壁仞、燧原等多家竞争。中国大模型有DeepSeek、Qwen、GLM、文心、混元等多家竞争。这种多元竞争格局，在一定程度上复制了市场竞争的创新激励。
- 多元竞争的另一种表现是A股高波上涨特征。
- **解法C："举国体制"集中攻关关键技术**
- 对于卡脖子的核心技术，中国采用举国体制集中资源攻关，用规模效应替代市场激励。

下载日志已记录，仅供内部参考

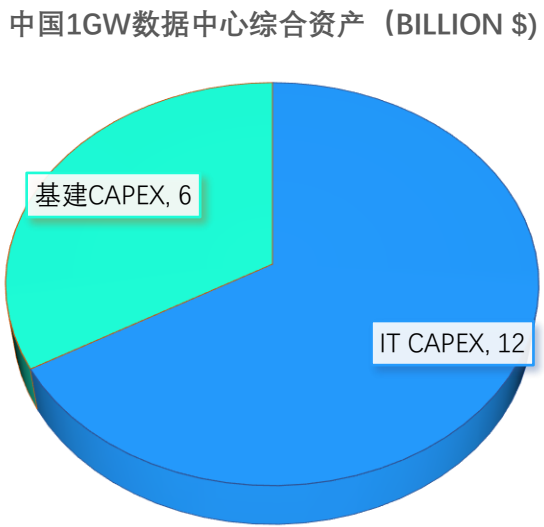
目录

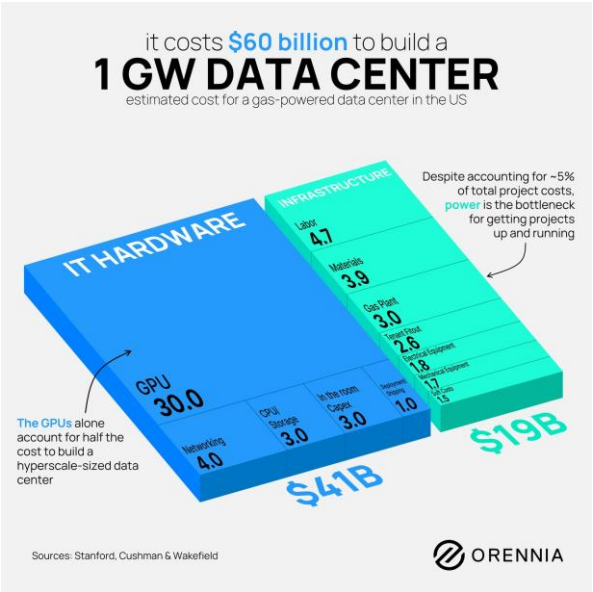
- AI即宏观：Token财政与美国经济新范式
- AI算力=制造：中美资本配置逻辑与竞争底层差异
- **中美1GW数据中心成本对比测算**
- 两条路径演化
- 风险提示

下载日志已记录，仅供内部参考

一座1GW数据中心，到底花了多少钱？

- **美国一个 1GW 燃气驱动的数据中心运行约 600 亿美元的综合资产。**IT 硬件和物理基础设施之间的比例约为 70/30。
- **中国1GW数据中心综合资产约180亿美元。**





资料来源：SemiAnalysis、AIJourn、Orennia、TrendForce，华福证券研究所

下载日志已记录，仅供内部参考

一座1GW数据中心，到底花了多少钱?

- 折旧怎么折?
- 数字硬件折旧年限不是假设，是三大公司10-K的明确披露
- 根据Epoch AI，基础设施加权平均折旧年限为14年

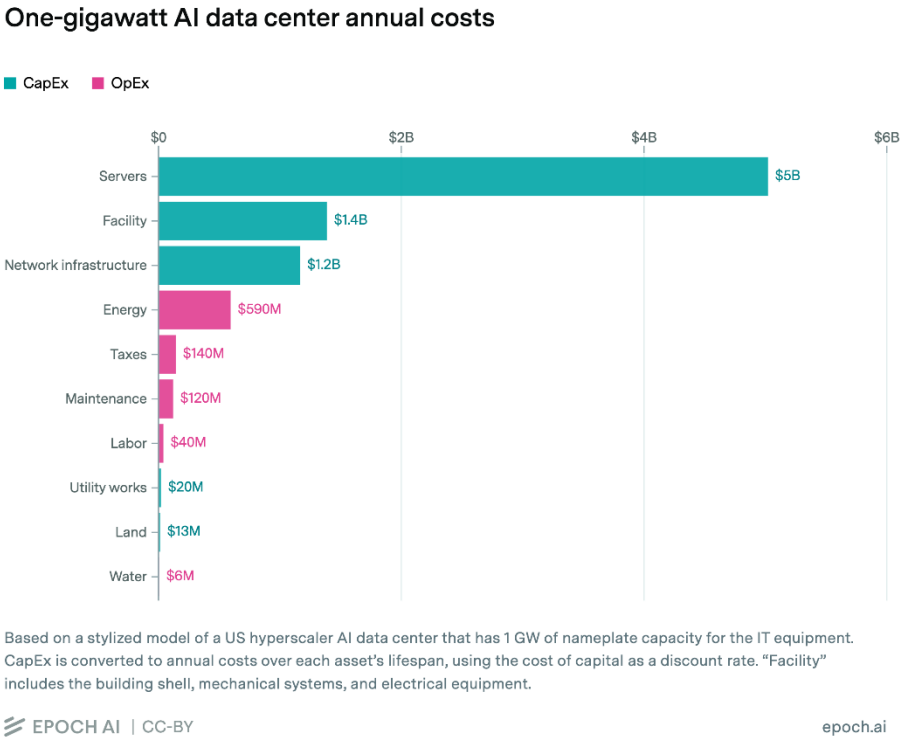
服务器/网络设备折旧年限（Amazon）	5年（2025年起）	"Effective January 1, 2025 we changed our estimate of the useful lives of a subset of our servers and networking equipment from six years to five years." 注：2024年曾从5年改为6年，2025年又改回5年。
服务器/网络设备折旧年限（Microsoft）	6年	"we determined we should increase the estimated useful lives of both server and network equipment from four years to six years. This change in accounting estimate was effective beginning fiscal year 2023."
服务器折旧年限（Alphabet/Google）	6年	"estimated useful life of our servers to six years. The effect of this change was a reduction in depreciation expense of $3.9 billion for the year ended [2023]."
训练侧IT硬件折旧年限	3年（<3年）	"the majority of AI spend are currently going toward IT hardware (mostly NVIDIA GPUs) with an incredibly short lifespan of **under three years** when used for training. This means that training requires a constant stream of reinvestment, making it more capitally intensive to do so on an ongoing basis."
基础设施加权平均折旧年限（本模型取值）	14年	" 5 years for IT equipment and 14 years for the facility."

资料来源：SEC、Amazon公告、Microsoft公告、Google公告、Alvarez & Marsal、Epoch AI，华福证券研究所

下载日志已记录，仅供内部参考

一座1GW数据中心，到底花了多少钱?

- **OpEx怎么算?**
- 根据Epoch AI，"$0.9 billion in annual operating expenses (OpEx)... even energy, the largest OpEx category, costs only $0.6 billion per year." （即电力约6亿/年，其余运维约3亿/年，**合计约9亿/GW/年**）

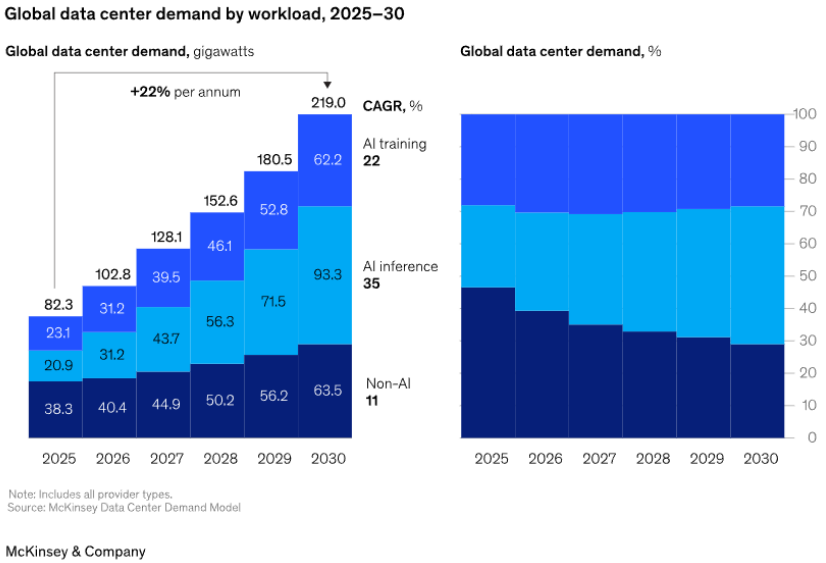


资料来源：Epoch AI，华福证券研究所

下载日志已记录，仅供内部参考

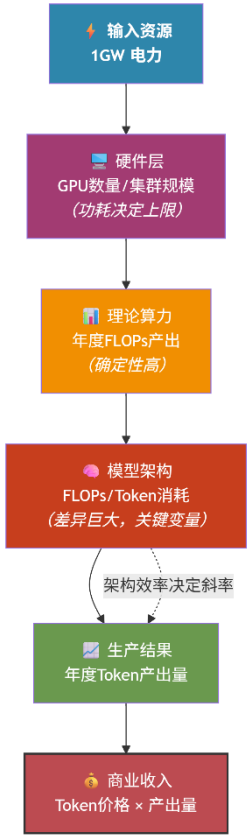
一座1GW数据中心，到底花了多少钱？

- **为了维持这1GW算力的商业竞争力，每年需要持续投入多少训练成本？**
  - 根据McKinsey，全球数据中心里，按需求来看，2026年推理和训练的比重接近。
- **我们假设：1GW数据中心里，50%的算力用于训练，50%用于推理。**训练侧的资本成本，按同等比例摊入。



资料来源：McKinsey，华福证券研究所

Token不是本质成本，FLOPs才是



- **三大测算难点**
- **难点1：GPU数量无法从功耗直接推算**
- 数据中心以"机柜产品"为单位部署（NVIDIA GB200 NVL72、华为CM384），每个机柜的功耗、卡数、价格是固定的，不能简单用"总功耗÷单卡功耗"来算。
- **难点2：生产实际吞吐量远低于FLOPs理论值**
- H100 FP16算力989 TFLOPS（密集），理论上可支持约2473 tok/s，但生产实际约200 tok/s，差了约12倍。Decode阶段受HBM内存带宽瓶颈制约，GPU大部分时间在等内存，不是在算。
- **难点3：中美硬件代际不同，口径难以统一**
- 美国Orennia数据基于H100时代，中国CM384是2025年产品。本测算设置五个场景覆盖不同代际，并单独列出"中国获得B200"的假设场景。

资料来源：华福证券研究所绘制

下载日志已记录，仅供内部参考

FLOPs理论产出 vs 生产实际——差距从哪里来

- **H100理论上能做2473 tok/s，生产实际只有200 tok/s——差了12倍**

度量维度	计算方式	H100数值	B200数值	差距（B200/H100）
FLOPs理论上限（tok/s/GPU）	FP16算力 ÷ 单token FLOPs消耗(GPT-4o 200B参数) H100：989 TFLOPS ÷ 400B FLOPs/tok B200：2250 TFLOPS ÷ 400B FLOPs/tok	~2,473 tok/s（989 TFLOPS ÷ 400B FLOPs/tok）	~5,625 tok/s（2,250 TFLOPS ÷ 400B FLOPs/tok）	2.27×
内存带宽单请求上限（tok/s/GPU，单请求）	HBM带宽 ÷ 模型权重大小 H100：3.35 TB/s ÷ （200B参数×FP16为2字节=400GB） B200：7.7 TB/s ÷ 400GB	~8 tok/s	~19.3 tok/s（7.7 TB/s ÷ 400GB，单请求）	2.30×（7.7÷3.35）
生产实际吞吐量（tok/s/GPU，batch优化）	实测数据（13B-70B模型，中等并发） Batch优化：多请求共享一次权重读取	~200 tok/s（实测）	~460 tok/s（推算：200×2.30）	2.30×（参考2.27×和2.30×）
FLOPs理论 vs 生产实际的差距	生产实际 ÷ FLOPs理论上限	200 ÷ 2,473 = 8%（FLOPs利用率约8%）	460 ÷ 5625 = 8%	—
Batch优化的效果（单请求→多请求并发）	多请求共享一次权重读取 有效带宽利用率提升	8 tok/s（单请求）→ 200 tok/s（batch） 提升约25倍	19.3 tok/s（单请求）→ 460 tok/s（batch） 提升约24倍	接近

注：单位B为Billion，单位GB为Gigabyte

资料来源：Nvidia官网、gmicloud、arxiv、macaron，华福证券研究所

下载日志已记录，仅供个人学习

解释

- **权重（weight）是什么**

- AI模型本质上是一堆数字（几百亿到几千亿个），存在显卡的内存里。这堆数字就是"模型学到的东西"，专业上叫"权重"。你可以把它想象成一本超级厚的参考书，每次模型要说一个字，都要把这本书从头到尾翻一遍，才能算出下一个字该是什么。这本"书"有多厚呢？以GPT-4o为例，激活参数200B（2000亿个数字），每个数字占2个字节，算下来这本书大概400GB。

- **内存带宽是什么、为什么会卡**

- 显卡上有两个东西：算力芯片：负责做数学运算，速度很快；显存（HBM）：负责存那本"书"，芯片要用的时候，得先把书从显存"搬"到芯片跟前"内存带宽"，就是显存往芯片搬东西的搬运速度。H100这张卡，搬运速度是每秒3.35TB。问题在这：模型每吐出一个字，都要把那本400GB的书完整搬一遍（因为每个字都要用到全部的权重）。搬一遍要多久？400GB ÷ 3.35TB/s ≈ 0.12秒也就是说，光是"把书搬到芯片跟前"这一步，就要花0.12秒，一秒钟最多搬8次多一点——这就是为什么单独伺候一个人聊天时，H100最多也就8个字/秒左右，算力芯片其实很闲，大部分时间在等书搬过来，不是在算。这就是"内存带宽瓶颈"：不是算得慢，是搬得慢，芯片天天在等快递。

- **Batch（批处理）是什么，为什么能解决一部分问题**

- 既然每次都要搬同一本书，那能不能"搬一次书，同时伺候好几个人"？比如同时有100个用户在跟这个AI聊天，与其每人单独搬一次书（100次搬运），不如把这100个人的请求凑在一起，搬一次书，100个人一起用——这就是batch（批处理）：把多个用户的请求打包在一起处理，摊薄"搬书"这个固定成本。这样效率一下就上去了：原来搬一次书只服务1个人，现在搬一次书服务100个人，虽然搬书的时间没变，但产出的字数翻了100倍。这也是为什么生产环境的实际速度（200字/秒）能比单人伺候时（8字/秒）快很多。

- **那为什么不把batch打包得更大，直接逼近理论上限?**

- 等人凑不齐会让人等太久：要凑够1000个人一起处理，可能要多等好几秒才能攒够人数，但用户发消息之后不可能接受等太久才开始回复，所以不能无限制地攒大批量，得在"效率"和"响应速度"之间找平衡。

- 显存不够装：每多打包一个人的请求，就要多占一块显存空间去记住"这个人聊到哪了"（这部分专业上叫KV cache，你可以理解成"每个人的聊天记录草稿纸"）。显存除了要装那本400GB的大书，还要给每个人留一份草稿纸，草稿纸攒多了，显存就装不下了，也没法再继续打包更多人。

下载日志已记录，仅供内部参考

中美模型FLOPs效率对比 + 信息不对称

- **中国模型用1/7的FLOPs生成同质量token**
- **开源 vs 闭源的信息不对称；中国模型完全透明 vs 美国模型完全不透明**：中国的成本优势是可量化、可验证的（开源数据）；美国的效率优势是不透明的（闭源）。在token价格竞争中，可量化的成本优势比不透明的技术优势更具说服力。这种信息不对称本身，就是中国开源策略的竞争优势之一。

模型	架构类型	总参数量	激活参数/token	单token FLOPs消耗（Decode近似）	可信度
GPT-4（2023）	稠密MoE（16 experts×111B）	1.8T	280B（2 experts×111B + 55B attention）	~560B FLOPs（≈2×280B）	★★★★☆
GPT-4o（估算）	优化MoE（比GPT-4更高效）	未公开	200B	~400B FLOPs（2×200B）	★★★☆☆
DeepSeek-V3（2024）	稀疏MoE（Multi-head Latent Attention）	671B	37B（官方技术报告直接披露）	~74B FLOPs（≈2×37B，Decode近似公式）	★★★★★
Qwen2.5-72B（2024）	密集模型（全参数激活）	72B	72B（密集模型，全部激活）	~144B FLOPs（≈2×72B）	★★★★★

资料来源： patmcguinness 、macaron 、arXiv ，华福证券研究所

下载日志已记录，仅供内部参考

五个场景的硬件参数与成本

- **五个场景：从H100基准到B200新时代，从华为CM384到假设获得NVIDIA**

场景	硬件	适用方	说明
A	H100	美国	Orennia基准，当前主流
B	B200	美国	新一代，2025年起部署
C	CM384（910C）	中国	华为现实方案，受产能限制
D1	H100	中国（假设）	出口管制解除情景
D2	B200	中国（假设）	出口管制解除情景

参数	A（美国H100）	B（美国B200）	C（中国CM384）	D1（中国H100）	D2（中国B200）
GPU/芯片型号	H100 SXM	HGX B200	昇腾910C（CM384）	H100 SXM	HGX B200
单卡/架功耗	700W/张	1,000W/张	600kW/架	700W/张	1,000W/张
GPU/芯片总数	92.3万张	90万张	57.6万片	92.3万张	90万张
IT硬件Capex	$410亿	$493亿	$120亿	$410亿	$493亿
基础设施Capex	$190亿	$190亿	$60亿	$60亿	$60亿
总Capex	$600亿	$683亿	$180亿	$470亿	$553亿

资料来源：SemiAnalysis、Nvidia、IntuitionLabs、Trendforce、Orennia、AIJourn，华福证券研究所

下载日志已记录，仅供内部参考

年度Token产出——五场景计算

- **HBM带宽决定B200的速度优势；CM384功耗过高是中国方案的核心劣势**

参数	场景A 美国H100	场景B 美国B200（HGX）	场景C 中国CM384（910C）	场景D1 中国H100（假设）	场景D2 中国B200（假设）
HBM内存带宽	3.35 TB/s	7.7 TB/s （HGX B200，官方）	3.2 TB/s （SemiAnalysis，非华为官方）	3.35 TB/s	7.7 TB/s （HGX B200，官方）
相对H100的带宽倍数	1.0×（基准）	2.30× （7.7÷3.35）	0.955× （3.2÷3.35）	1.0×	2.30×
生产实际吞吐量 （tok/s/GPU，中等模型）	200 tok/s （实测）	460 tok/s （推算：200×2.30） ⚠推算值	~191 tok/s （推算：200×0.955） ⚠推算值	200 tok/s （实测）	460 tok/s （推算） ⚠推算值
GPU/芯片总数	92.3万张	90万张	57.6万片 （1,500架×384片）	92.3万张	90万张
年有效运行时间（秒）	2.52×10^7s （80%利用率）	2.52×10^7s （80%利用率）	2.52×10^7s （80%利用率）	2.52×10^7s （80%利用率）	2.52×10^7s （80%利用率）
年度Token产出量	**4.65×10^{15}/年 （92.3万 ×200×2.52×10^7）**	**10.43×10^{15}/年 （90万×460×2.52×10^7）**	**2.77×10^{15}/年 （57.6万×191×2.52×10^7）**	**4.65×10^{15}/年**	**10.43×10^{15}/年**
相对场景A（美国H100）的倍数	1.0×（基准）	2.24× （10.43÷4.65）	0.6× （2.77÷4.65）	1.0×	2.24×

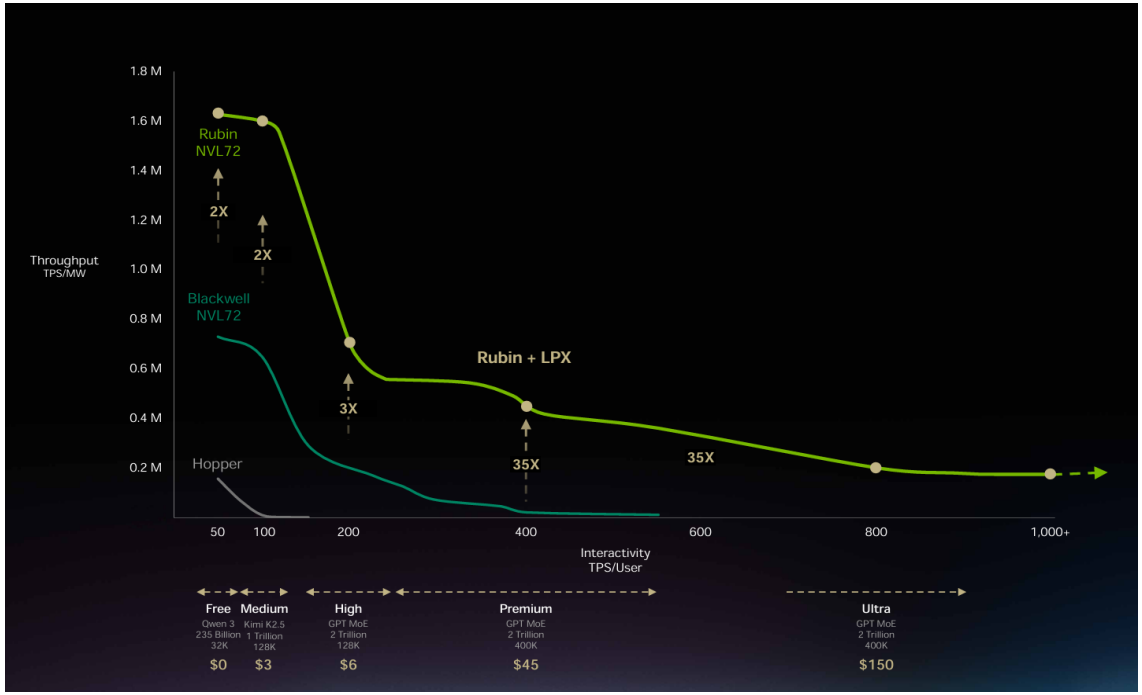
注：①模型参数：适用于13B-70B模型，大模型实际更低②利用率：乐观情景（70%-80%，采用80%）

资料来源：SemiAnalysis、Nvidia、gmicloud、Orennia，华福证券研究所

下载日志已记录，仅供内部参考

模型越复杂，同样的算力产出的token越少——我们的测算是保守估算

- **横轴：每用户每秒token速率（TPS/User），代表模型的"交互性"要求——用户要求的响应速度越快，这个数字越大**
- **纵轴：每兆瓦的总吞吐量（TPS/MW），代表同样的电力预算下，能产出多少token**



H卡在横轴约50 TPS/用户时，纵轴约200,000 TPS/MW。

换算成每张GPU：
H100功耗约700W = 0.0007 MW
每张GPU吞吐 = 200,000 TPS/MW × 0.0007 MW = 140 tok/s/GPU

这和我们用的200 tok/s是同一个量级

资料来源：Nvidia，华福证券研究所

下载日志已记录，仅供内部参考

为什么会这样——背后的物理原理

- **模型越大，每个token消耗的算力越多，同样的电力能产出的token越少。**
- 具体来说，对MoE架构的模型：
  - 每次生成一个token，GPU需要读取"激活的专家参数"的全部权重
  - 激活的专家参数越多（模型越大），需要读取的数据量越大
  - HBM内存带宽是固定的，读取时间更长，token速率自然更低
- 类比：同样一辆卡车（GPU），装的货越重（模型越大），跑得越慢，单位时间运的货越少。
- **两个维度叠加：**
  - **模型越大 → 每个token消耗的算力越多 → 吞吐量越低**
  - **用户要求响应越快 → batch size越小 → GPU利用率越低 → 吞吐量越低**

下载日志已记录，仅供内部参考

我们的测算在哪里——为什么是保守的

- **我们的测算假设：H100，200 tok/s，对应13B-70B中等模型，中等并发**
- 在NVIDIA这张图上，这对应的是：
  - 横轴约100-200 TPS/用户（Medium到High之间）
  - 纵轴约200,000 TPS/MW（Hopper曲线）
  - 换算单卡：200,000 TPS/MW × 0.0007 MW/卡（H100功耗700W）= 140 tok/s/卡
- **与我们用的200 tok/s基本一致。**
- **为什么说我们的测算是保守的?**
  - 美国CSP实际部署的主力商业模型（如GPT MoE）属于图中"High到Premium"区间
  - 用户要求响应速度更快，模型规模总参数和激活参数大概率更大
  - 对应Hopper架构的吞吐量更小
- **但由于美国模型（如GPT-5、Claude Opus 4）完全闭源，我们无法获得准确的激活参数量，因此无法精确计算其吞吐量。我们选择用中等模型（200 tok/s）作为统一基准，是保守且可验证的假设。**

下载日志已记录，仅供内部参考

年度固定成本汇总

- **中国基础设施和融资成本优势明显**

成本项	场景A 美国H100	场景B 美国B200	场景C 中国CM384	场景D1 中国H100	场景D2 中国B200
IT硬件Capex（亿美元）	410	493	120	410	493
基础设施Capex（亿美元）	190	190	60	60	60
总Capex（亿美元）	600	683	180	470	553
推理侧IT折旧 （50%×IT÷5年，亿美元/年）	41	49.3	12	41	49.3
训练侧IT折旧 （50%×IT÷3年，亿美元/年）	68.3	82.2	20	68.3	82.2
基础设施折旧 （÷14年，亿美元/年）	13.6	13.6	4.3	4.3	4.3
纯运营OpEx （电费/人工/维保，亿美元/年）	9	9	9	9	9
年固定成本合计（亿美元/年）	131.9	154.1	45.3	122.6	144.8
WACC	8%	8%	5%	5%	5%
税率	21%	21%	25%	25%	25%

注：该表格仅为示意性测算

资料来源：AIJourn、Epoch AI、McKinsey、人民网、Britannica、 Convequity、 ifind、国家税务总局、Aswath Damodaran 、Orennia，华福证券研究所

下载日志已记录，仅供内部参考

盈亏反算——五场景核心结论

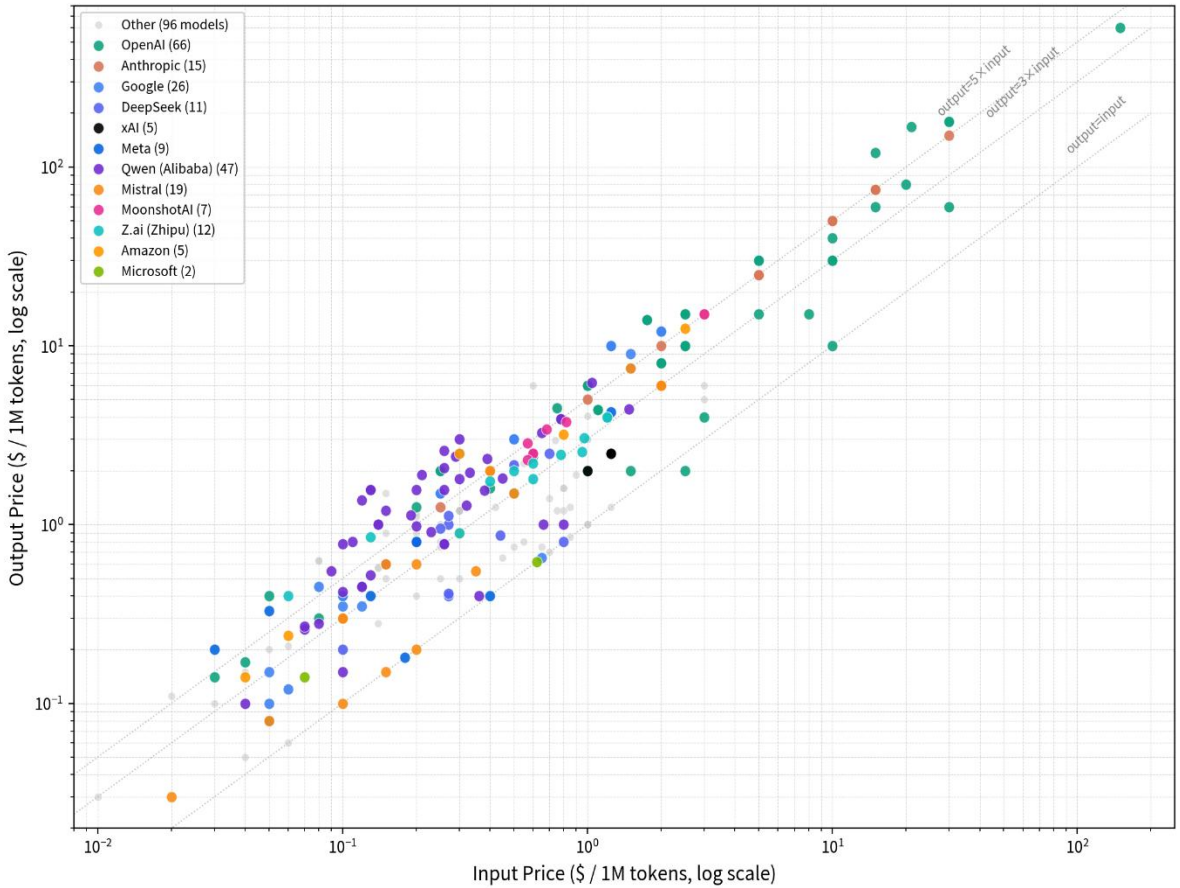
- **反算公式：盈亏平衡 = 年固定成本 ÷ (1-25%可变OpEx比率) ÷ 年Token产出量**
- **覆盖WACC：目标NOPAT = WACC×IC，目标营业利润 = NOPAT÷(1-税率)，所需收入 = (营业利润+固定成本)÷75%，Token价格 = 收入÷产出量**

指标	场景A 美国H100	场景B 美国B200	场景C 中国CM384	场景D1 中国H100	场景D2 中国B200
年固定成本（亿美元/年）	131.9	154.1	45.3	122.6	144.8
年Token产出量（×10^{15}）	4.65	10.43	2.77	4.65	10.43
总Capex（亿美元）	600	683	180	470	553
WACC	8%	8%	5%	5%	5%
税率	21%	21%	25%	25%	25%
盈亏平衡Token价格 （美元/百万token）	$3.78/M	$1.97/M	$2.18/M	$3.52/M	$1.85/M
覆盖WACC所需Token价格	$5.52/M （WACC=8%）	$2.85/M （WACC=8%）	$2.76/M （WACC=5%）	$4.41/M （WACC=5%）	$2.32/M （WACC=5%）
达到CSP ROIC所需Token价格	$7.95/M （目标19%）	$4.08/M （目标19%）	$3.68/M （目标13%）	$5.85/M （目标13%）	$3.08/M （目标13%）

注：该表格仅为示意性测算。可变OpEx比率25%为情景测算简化假设，用以刻画随收入变动的运营成本（如带宽/客服/云服务分成等），未采用云厂商分部披露数据逐项拆分核算。

资料来源：SemiAnalysis、Nvidia官网、gmicloud、Orennia、AIJourn、SEC、Amazon公告、Alvarez & Marsal、Epoch AI、McKinsey、IntuitionLabs、Trendforce、Convequity、ifind、人民网、国家税务总局、Aswath Damodaran、britannica、Findbox、Stock Analysis on Net，华福证券研究所

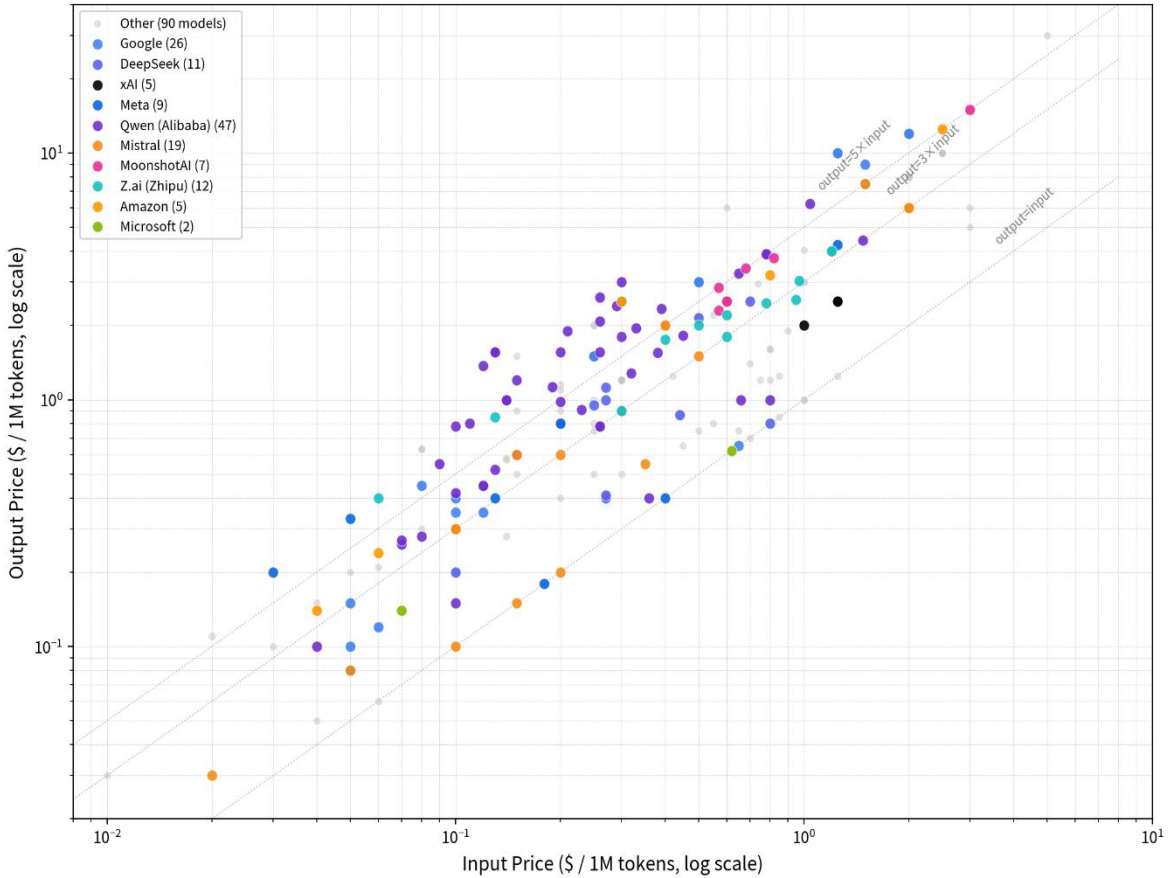
Token市场价格



资料来源：Iternal，华福证券研究所

下载日志已记录，仅供内部参考

Token市场价格（剔除OpenAI和Anthropic）



资料来源：Iternal，华福证券研究所

下载日志已记录，仅供内部参考

目录

- AI即宏观：Token财政与美国经济新范式
- AI算力=制造：中美资本配置逻辑与竞争底层差异
- 中美1GW数据中心成本对比测算
- **普惠供给与全球竞争重估**
- 风险提示

下载日志已记录，仅供内部参考

“以人为本、全球共用”，中国AI供给节奏是关键锚点

- **AI算力不是商品，是基础设施；不是竞争工具，是全球公共事业。这一定性直接挑战美国科技巨头的商业模式前提。**

> "人工智能是全人类共同的智慧结晶和宝贵财富……人工智能发展不应该是某个国家的独奏，而应当是全球合作的交响。"
>
> ——习总书记，2026世界人工智能大会主旨讲话，2026年7月17日，上海

习总书记提出的原则	对美国科技巨头的冲击含义
开放共赢：鼓励开源开放、合作共享	DeepSeek开源模式 vs OpenAI/Anthropic闭源收费模式，开源压制闭源溢价
以人为本、向上向善：让AI惠及千行百业	普惠定价（低token价格）成为方向，高价算力租赁模式面临压力
弥合数智鸿沟：帮助全球南方加强能力建设	中国以低成本算力+开源模型向发展中国家输出，绕开美国CSP的全球市场
反对泛化国家安全概念：反对把本国安全凌驾于他国安全之上	直接回应美国芯片出口管制，为中国算力出海提供政治话语支撑

资料来源：中国政府网、新华社，华福证券研究所

下载日志已记录，仅供内部参考

节奏锚定：中国AI供给增速是全球token价格的天花板

- **当AI被定位为全人类共同的宝贵财富、中国致力于做全球公共产品的持续提供者时，算力供给逻辑或将从紧平衡稀缺转向留有余量的冗余建设，这也是普惠承诺得以持续兑现的重要基础之一。中国的战略资本就不再是"补贴"，而是基础设施投资的天然属性；美国的金融资本回报逻辑，在这一框架下面临根本性重估。**
- **中国定价目标是普惠：**习总书记明确“普惠如何实现”是时代之问，AI算力首先应服务于人，中国AI供给以压低token价格为政策导向，而非追求利润最大化。
- **中国供给增速决定全球价格下限：**中国以战略资本驱动算力扩张，成本结构天然低于美国；只要中国算力持续扩张，全球token价格就有持续下行压力。
- **美国CSP的盈亏平衡被动锚定在中国供给节奏上：**只要中国以低于美国盈亏平衡的Token价格供给市场，美国算力投资的ROIC逻辑就无法成立。

下载日志已记录，仅供内部参考

两条演化路径

- **中国成本优势，是美国AI Capex叙事的系统性风险**
- 中国token价格持续下行 → 全球token价格天花板被压低 → 美国CSP收入预期下修 → ROIC进一步恶化→ AI Capex扩张的财务逻辑动摇
- **结局取决于一个问题：芯片封锁能否持续有效?**
- **路径A：美国维持AI垄断（芯片封锁有效）**
- 中国只能用华为硬件，绝对性能落后
- 但成本优势仍在，中低端推理市场被中国占据
- 美国CSP保住高端市场，但全球token价格天花板被中国压低
- **路径B：AI技术扩散（芯片封锁失效或华为追赶）**
- 中国在成本和性能上同时具备竞争力
- 全球token价格被系统性压低
- 美国CSP的ROIC进一步恶化，AI Capex扩张的财务逻辑动摇
- 类似"中国制造"的历史：先用成本优势占领市场，再提升技术

下载日志已记录，仅供内部参考

目录

- AI即宏观：Token财政与美国经济新范式
- AI算力=制造：中美资本配置逻辑与竞争底层差异
- 中美1GW数据中心成本对比测算
- 普惠供给与全球竞争重估
- 风险提示

下载日志已记录，仅供内部参考

风险提示

- **美国AI应用与生产率兑现超预期**
- 若AI渗透与生产率提升快于预期，CSP现金流内生改善、债务扩张转为可持续，本报告Token财政框架可能被证伪。
- **Token价格提前见顶回落**
- 若需求不及预期或竞争加剧压低价格，CSP收入与资本开支快速下修，触发信用收缩，工业通胀接棒判断不成立。
- **AI支出向传统工业品转移不及预期**
- 若电网接入、设备交付与施工许可瓶颈超预期，或CSP推迟基建投资，工业品需求兑现时点后移。
- **成本测算关键假设存在不确定性**
- 吞吐量、利用率、折旧与资本成本均为情景假设；美国模型闭源，效率无法验证。若其推理优化或硬件迭代优于假设，中美成本差距可能收窄。
- **国产算力进程与中美产业政策超预期变化**
- 若先进制程与关键配套持续受限，国产能效差距或长期存在；若出口管制收紧或对华加征关税，将冲击成本路径与对美出口链条。

下载日志已记录，仅供内部参考

分析师声明及一般声明

分析师声明

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

一般声明

华福证券股份有限公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告的信息均来源于本公司认为可信的公开资料，该等公开资料的准确性及完整性由其发布者负责，本公司及其研究人员对该等信息不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，之后可能会随情况的变化而调整。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

在任何情况下，本报告所载的信息或所做出的任何建议、意见及推测并不构成所述证券买卖的出价或询价，也不构成对所述金融产品、产品发行或管理人作出任何形式的保证。在任何情况下，本公司仅承诺以勤勉的职业态度，独立、客观地出具本报告以供投资者参考，但不就本报告中的任何内容对任何投资做出任何形式的承诺或担保。投资者应自行决策，自担投资风险。

本报告版权归“华福证券股份有限公司”所有。本公司对本报告保留一切权利。除非另有书面显示，否则本报告中的所有材料的版权均属本公司。未经本公司事先书面授权，本报告的任何部分均不得以任何方式制作任何形式的拷贝、复印件或复制品，或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。未经授权的转载，本公司不承担任何转载责任。

下载日志已记录，仅供内部参考

特别声明及投资声明评级

特别声明

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。投资者请勿将本报告视为投资或其他决定的唯一参考依据。

投资评级声明

类别	评级	评级说明
公司评级	买入	未来6个月内，个股相对市场基准指数涨幅在20%以上
	持有	未来6个月内，个股相对市场基准指数涨幅介于10%与20%之间
	中性	未来6个月内，个股相对市场基准指数涨幅介于-10%与10%之间
	回避	未来6个月内，个股相对市场基准指数涨幅介于-20%与-10%之间
	卖出	未来6个月内，个股相对市场基准指数涨幅在-20%以下
行业评级	强于大市	未来6个月内，行业整体回报高于市场基准指数5%以上
	跟随大市	未来6个月内，行业整体回报介于市场基准指数-5%与 5%之间
	弱于大市	未来6个月内，行业整体回报低于市场基准指数-5%以下

备注：评级标准为报告发布日后的6~12个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中，A股市场以沪深300指数为基准；香港市场以恒生指数为基准；美股市场以标普500指数或纳斯达克综合指数为基准（另有说明的除外）。

下载日志已记录，仅供内部参考

华福证券
HUAFU SECURITIES CO.,LTD.

诚信专业 发现价值

联系方式

华福证券研究所 上海
公司地址：上海市浦东新区浦明路1436号陆家嘴滨江中心MT幢20层
邮编：200120
邮箱：hfyjs@hfzq.com.cn