仅供内部参考，请勿外传

从电力到 Token：AI 算力的能源经济学

煤炭

评级： 增持

姓名	电话	邮箱	登记编号
黄涛(分析师)	021-38674879	huangtao@gtht.com	S0880515090001

本报告导读：

我们从能源视角重新构建 AI 算力成本研究框架，沿着“电力—算力—Token”的完整链条，将能源供需研究由电力需求进一步延伸至 AI 单位服务成本与长期竞争力。

投资要点：

- 我们从能源视角重新构建 AI 算力成本研究框架，沿着“电力—算力—Token”的完整链条，将能源供需研究由电力需求进一步延伸至 AI 单位服务成本与长期竞争力。我们认为当前 AI 成本仍以硬件投入为主，但随着芯片性价比、系统效率和利用率持续提升，稳定、低成本电力在单位有效算力及 Token 成本中的重要性有望持续上升，并最终成为影响 AI 长期竞争力的重要变量。我们继续强调全面看多未来 5-10 年全球能源大周期，煤炭我们继续推荐兖煤澳洲、兖矿能源、中国神华、陕西煤业、中煤能源、晋控煤业；受益标的 BTU、力量发展。
- 从电力到算力：AI 扩张正受到完整物理基础设施的约束。我们认为，AI 产业研究需要从模型和芯片进一步延伸至电力与数据中心。训练形成并更新模型能力，推理持续向用户交付服务，两者都需要由芯片、服务器、网络及冷却系统共同承载，并转化为真实的电力消耗。随着 AI 负荷呈现规模大、扩张快、连续运行要求高的特征，算力建设与电力配套之间的周期错配更加突出。因此，拥有 GPU 并不等于能够及时形成有效算力，电网接入、供电容量与散热能力，正在共同决定 AI 基础设施的落地速度和扩张空间。
- AI 成本竞争：短期看硬件，实际看系统效率，中长期看能源条件。我们认为，AI 算力经济性应从单卡采购价格与单卡性能扩展至数据中心全生命周期成本，并结合实际产出进行评价。当前 GPU 及服务器仍是成本主体，芯片性能与采购价格决定成本基础；但理论算力能否转化为有效产出，取决于互联架构、软件效率、任务调度及设备利用率。能源从短期角度已经成为部分限制 AI 发展的瓶颈所在，而电价和 PUE 持续影响运行成本。当硬件与系统效率逐步进入可比区间，稳定、低成本且能够持续扩张的电力供给，有望成为区域 AI 成本竞争中更难复制的优势。
- 从数据中心成本到 Token 经济性：我们认为，随着 AI 应用加快普及，产业研究需要进一步从算力和数据中心投入，延伸至服务产出与商业回报。推理将模型能力持续转化为面向用户的服务，Token 由此成为连接使用规模、商业定价和底层成本的重要计量单位；在模型质量与服务要求可比的条件下，将推理业务对应的基础设施成本与实际 Token 产出相匹配，可以进一步识别硬件、系统效率、设备利用率及能源条件对单位服务成本的影响。由此，AI 竞争在持续追求模型能力突破的同时，也将更加重视商业化交付的经济性：先进算力决定能力边界，系统效率影响成本与产出，稳定且低成本的能源供给则支撑持续运营与规模扩张。能否将算力投入转化为具有成本竞争力的服务，并形成可持续的商业回报，将成为衡量 AI 企业长期竞争力的重要标准。
- 风险提示：AI 应用及推理需求增长低于预期，AI 软硬件效率提升超预期。

相关报告

煤炭《煤价如期突破千元关口，关注冬储兑现及三季报业绩改善》2026.10.11
煤炭《国泰海通煤炭行业基本面数据库大全 20261010》2026.10.10
煤炭《冬储预期渐浓，煤价阶段性支撑仍在》2026.10.07
煤炭《国泰海通煤炭行业基本面数据库大全 20261007》2026.10.07
煤炭《前瞻预判如期兑现，冬储提供阶段性支撑》2026.09.29

仅供内部参考，请勿外传

目录

1. 投资概要 3
2. AI 从“算力竞争”到“基础设施竞争”，物理约束开始显现 4
2.1. AI 算力扩张进入物理基础设施约束阶段 4
2.2. AI 负荷呈现“大、快、刚”三大特征，传统电力系统难以同步扩张 5
2.2.1. “大”：数据中心从 MW 级走向 GW 级，单体负荷正在城市化 5
2.2.2. “快”：AI 基础设施建设周期明显短于电力基础设施 6
2.2.3. “刚”：实时 AI 服务对连续性要求高，负荷调节存在经济与技术约束 6
3. AI 耗电拆解：从模型计算到数据中心电力消耗 7
3.1. 从算法到算力：AI 运行依赖真实的物理载体 8
3.2. 训练与推理：AI 算力的两种基本形态 8
3.3. 从算力到数据中心：AI 计算最终落到实际的能源电力消耗 9
4. AI 成本竞争，短期看硬件，实际看系统利用率，中长期看能源 11
4.1. AI 数据中心成本结构：硬件是核心，运营成本中电力能源是大头 11
4.2. 短期看硬件：当前 AI 算力成本仍由芯片与服务器主导 13
4.3. 实际看效率：理论算力并不等于真实产出 14
4.4. 能源影响扩张能力与区域竞争力，硬件成本被持续摊薄后，能源的重要性将逐步上升 15
5. 从数据中心成本到 AI 商业化 17
5.1. AI 商业模式推进，token 成为经济性计量单位 18
5.2. 从数据中心 TCO 到 Token Cost：AI 成本竞争开始具备统一比较口径 19
5.3. 从算力规模到 Token 经济性：AI 竞争重心或将逐步由能力向成本迁移 20
6. 风险提示 23

仅供内部参考，请勿外传

1. 投资概要

"Electricity powers intelligence. The economics of turning it into useful tokens shapes the value of AI."

在此前的全球能源系列研究中，我们从全球缺电出发，探讨电力需求增长、供给结构变化与传统能源的系统价值，AI 是其中重要的需求变量。我们认为，随着 AI 应用走向规模化，理解其竞争力需要同时考察模型能力与背后的资源投入、生产效率和成本结构。从电力到 Token，正是将能源供需研究延伸至 AI 单位服务经济性的完整链条，也是我们理解能源禀赋如何转化为数字产业竞争优势的新视角。而这也是我们从全球能源研究进一步进入 AI 成本和 Token 经济学研究的起点。

从电力到算力：AI 扩张正受到完整物理基础设施的约束。我们认为，AI 产业研究需要从模型和芯片进一步延伸至电力与数据中心。训练形成并更新模型能力，推理持续向用户交付服务，两者都需要由芯片、服务器、网络及冷却系统共同承载，并转化为真实的电力消耗。随着 AI 负荷呈现规模大、扩张快、连续运行要求高的特征，算力建设与电力配套之间的周期错配更加突出。因此，拥有 GPU 并不等于能够及时形成有效算力，电网接入、供电容量与散热能力，正在共同决定 AI 基础设施的落地速度和扩张空间。

AI 成本竞争：短期看硬件，实际看系统效率，中长期看能源条件。我们认为，AI 算力经济性应从单卡采购价格与单卡性能扩展至数据中心全生命周期成本，并结合实际产出进行评价。当前 GPU 及服务器仍是成本主体，芯片性能与采购价格决定成本基础；但理论算力能否转化为有效产出，取决于互联架构、软件效率、任务调度及设备利用率。能源从短期角度已经成为部分限制 AI 发展的瓶颈所在，而电价和 PUE 持续影响运行成本。当硬件与系统效率逐步进入可比区间，稳定、低成本且能够持续扩张的电力供给，有望成为区域 AI 成本竞争中更难复制的优势。

从数据中心成本到 Token 经济性：我们认为，随着 AI 应用加快普及，产业研究需要进一步从算力和数据中心投入，延伸至服务产出与商业回报。推理将模型能力持续转化为面向用户的服务，Token 由此成为连接使用规模、商业定价和底层成本的重要计量单位；在模型质量与服务要求可比的条件下，将推理业务对应的基础设施成本与实际 Token 产出相匹配，可以进一步识别硬件、系统效率、设备利用率及能源条件对单位服务成本的影响。由此，AI 竞争在持续追求模型能力突破的同时，也将更加重视商业化交付的经济性：先进算力决定能力边界，系统效率影响成本与产出，稳定且低成本的能源供给则支撑持续运营与规模扩张。能否将算力投入转化为具有成本竞争力的服务，并形成可持续的商业回报，将成为衡量 AI 企业长期竞争力的重要标准。

表1：重点覆盖公司估值表

公司简称	代码	eps（元）			pe			投资评级	收盘价（元）
		2024A	2025A	2026E	2024A	2025A	2026E		
新集能源	601918.SH	0.92	0.83	0.99	11.6	13.0	10.8	增持	10.70
晋控煤业	601001.SH	1.68	1.09	1.11	9.9	15.3	15.0	增持	16.64
电投能源	002128.SZ	2.38	2.42	3.32	11.6	11.4	8.3	增持	27.53
恒源煤电	600971.SH	0.89	(0.16)	0.43	9.2	-	19.0	增持	8.17
中国神华	601088.SH	2.95	2.66	3.08	16.7	18.6	16.0	增持	49.38
陕西煤业	601225.SH	2.31	1.73	2.45	11.8	15.8	11.2	增持	27.35
华阳股份	600348.SH	0.59	0.43	0.53	15.0	20.6	16.7	增持	8.87
兖矿能源	600188.SH	1.46	0.84	1.52	13.5	23.5	13.0	增持	19.71

仅供内部参考，请勿外传

中煤能源	601898.SH	1.46	1.35	1.41	10.0	10.8	10.3	增持	14.59
兰花科创	600123.SH	0.48	(0.37)	0.20	14.0	-	33.9	增持	6.77
山煤国际	600546.SH	1.14	0.59	1.12	11.8	22.9	12.1	增持	13.50
山西焦煤	000983.SZ	0.55	0.21	0.46	11.8	30.6	14.1	增持	6.48
平煤股份	601666.SH	0.96	0.09	0.67	9.1	93.7	13.1	增持	8.76
淮北矿业	600985.SH	1.84	0.56	1.37	8.6	28.3	11.5	增持	15.82
潞安环能	601699.SH	0.82	0.37	1.14	18.9	41.8	13.6	增持	15.47
兖煤澳大利亚	3668.HK	4.49	1.73	2.38	7.1	18.5	13.4	增持	31.92

数据来源：Choice，国泰海通证券研究（注：收盘价截至 2026/10/12）

2. AI 从“算力竞争”到“基础设施竞争”，物理约束开始显现

我们认为，随着 AI 从模型训练逐步进入大规模商业化和持续推理阶段，决定 AI 产业扩张速度的瓶颈正在从单一芯片供给，逐步向整套物理基础设施扩散，能源和基础设施正在成为决定 AI 算力扩张速度、成本和最终竞争力的核心生产要素。

2.1. AI 算力扩张进入物理基础设施约束阶段

我们看到 2023 年开始资本市场对于人工智能产业的研究核心更多集中在模型能力和算力供给。随着大语言模型快速迭代，参数规模、训练算力、GPU 数量、芯片性能以及先进制程成为衡量 AI 产业发展的核心指标，AI 竞争也一度被简化为“谁能够获得更多且最先进的英伟达的卡”。这一框架在 AI 产业发展的早期阶段具有很强的解释力：模型规模快速扩大，训练所需计算量持续提升，GPU 成为最稀缺的生产资料，算力供给直接决定模型迭代速度和能力上限。

随着 AI 训练规模持续扩大、推理应用加快普及，算力扩张对配套基础设施的要求也在提高。GPU 只有完成部署，并获得稳定供电、散热和网络互联支持，才能转化为可持续运行的有效算力。因此，研究 AI 供给能力需要同时考察芯片获取与基础设施交付：即使企业能够采购足够的 GPU，电网接入、变电容量或冷却系统建设滞后，仍可能推迟算力投运。电力与数据中心基础设施由此成为影响 AI 扩张速度的重要约束。这也是为什么从 2025H2 开始“美国缺电”问题真正被市场认知到的核心原因，**电力系统作为物理的整套基础设施正在成为制约 AI 发展的另一重要瓶颈。**

从产业链角度看，一次 AI 计算任务背后实际上对应一条完整的物理链条：电力首先进入数据中心，经由变电、UPS 和配电系统输送至服务器，再由 GPU、CPU、HBM、网络交换设备和存储系统完成计算，同时制冷系统持续带走计算产生的热量，最终形成可以被模型调用的有效算力。其中任何一个环节出现瓶颈，都可能限制最终算力的释放。拥有 GPU 并不等于拥有可用算力，拥有理论算力也不等于能够稳定、低成本地向用户提供 AI 服务。

这一变化在 AI 负荷快速扩张后变得尤为明显。大型 AI 数据中心正在从过去几十 MW 的传统数据中心，快速向数百 MW 乃至 GW 级负荷演进，而数据中心的建设和扩张速度明显快于电网、电源及部分关键设备的建设周期。我们在此前全球能源系列中已经反复讨论，AI 数据中心正在成为全球新增电力需求的重要边际变量，并对电网接入、可靠容量和稳定电源提出更高要求。

仅供内部参考，请勿外传

图1：AI 算力背后是完整的物理基础设施体系



资料来源：国泰海通证券研究

2.2. AI 负荷呈现“大、快、刚”三大特征，传统电力系统难以同步扩张

我们认为，相较居民、商业乃至大多数工业负荷，AI 数据中心最显著的特征在于**单体规模更大、负荷落地更快、用电需求更刚性**。这三种特征叠加，使 AI 对电力系统造成的压力远高于其单纯的年度用电量占比，也决定了未来研究 AI 不能只聚焦“电量”概念，更需要关注电力基础设施“负荷容量”与提供的速度。

2.2.1. “大”：数据中心从 MW 级走向 GW 级，单体负荷正在城市化

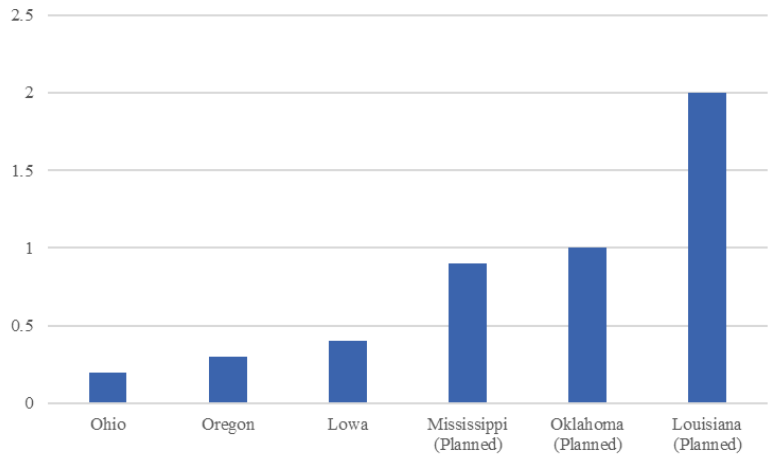
我们认为 AI 数据中心正在把过去分散增长的电力需求，转变为高度集中的 GW 级“超级负荷”。传统数据中心的单体负荷多数处于几十 MW 至数百 MW 区间，但随着大模型训练规模持续扩大以及 AI 推理需求快速增长，新建 AI 数据中心正明显向 GW 级演进。

我们此前在全球能源系列中已经观察到，部分头部企业规划中的大型 AI 数据中心园区正向 GW 级扩张。与此同时，数据中心往往集中布局于少数区域，使新增负荷在有限的地理范围内快速接入。即使全国全年发电量能够覆盖用电需求，也不代表特定地区在高峰时段拥有充足的供电能力；局部输电通道、变电容量及可靠电源不足，均可能限制项目接入。因此，评估 AI 电力供需，需要同时关注全年电量、峰值功率以及负荷所在区域的电网承载能力。

因此，AI 对电力系统的第一个改变是显著提高了**负荷的空间集中度**，导致未来电力短缺越来越可能表现为局部节点和特定区域的容量不足，而非传统意义上的全国性电量短缺。

仅供内部参考，请勿外传

图2：未来新建 AI 数据中心耗电都是 GW 级负荷消耗（GW）



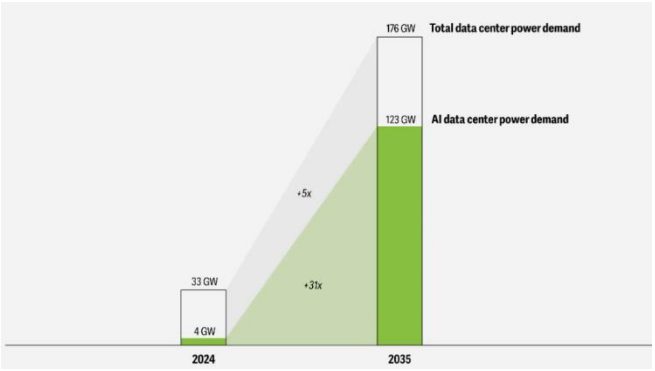
资料来源：Deloitte，国泰海通证券研究

2.2.2. “快”：AI 基础设施建设周期明显短于电力基础设施

我们认为，AI 负荷的第二个特征在于扩张速度与电力配套建设周期之间的错配。IEA 指出，数据中心可以在约 2—3 年内投运，而配套能源基础设施通常需要更长的规划和建设时间。具体项目的周期取决于土地、许可、设备交付及电网接入条件，数据中心施工完成也不代表能够立即获得足额供电。因此，判断项目落地速度，应区分机房建设、设备部署与电力接入三个环节，重点关注其中交付最慢的约束。

AI 产业的投资计划随模型进展、应用需求及竞争格局快速调整，而电网和电源建设需要结合长期负荷预测、审批及工程实施推进。当科技企业上调算力建设计划时，原有电力规划未必能够同步响应，容易形成“算力投资先行、电力配套滞后”的局面。由此，能够提前落实供电容量和接入条件的项目，更有可能缩短设备等待时间，并率先将资本投入转化为实际算力产出。

图3：德勤预计美国 2035 年 AI 数据中心用电达到 123GW，相对 2024 年翻 30 倍



资料来源：Deloitte Research Center for Energy & Industrials

图4：AI 数据中心建设周期平均在 2 年左右，非常快速

Timeline for data center construction		
Phase	Duration	What's Included
Planning & Feasibility	3-6 months	Site review, demand analysis, capex modeling, grid checks
Design & Engineering	6-12 months	Concept layouts, Tier strategy, cooling and power design
Permits & Approvals	6-18 months (overlaps)	Zoning, environmental, utility approvals
Construction	12-24 months	Civil works, MEP build, shell, and system installation
Testing & Commissioning	3-6 months	Load tests, system integration, performance certifications

资料来源：Mastt

2.2.3. “刚”：实时 AI 服务对连续性要求高，负荷调节存在经济与技术约束

我们认为，AI 负荷的第三个特征，在于较高的服务连续性要求和固定资产利用需求。对于在线对话、搜索及企业实时应用而言，服务中断会直接影响用户体验与业务运行，因此相关推理负荷的短期调节空间相对有限。但不同

仅供内部参考，请勿外传

AI 任务的灵活性存在差异：训练、离线推理及部分可延迟任务，可以在满足交付要求的前提下调整运行时间；具备多地部署能力的企业，还可以通过跨区域调度分散负荷。AI 用电具有较强刚性，但其程度取决于任务类型和系统设计。

从经济性看，GPU、服务器及机房等固定资产投入较高，企业具有提高利用率、保持持续运行的动力。在本文引用的 Epoch AI 典型数据中心模型中，电费约占年度总成本的 7%，因此，在常规电价波动下，为节省电费而大幅降低设备利用率，未必能够改善整体经济性。不过，当尖峰电价、需求响应补偿或提前接电的收益足够高时，适度调整负荷也可能具有经济价值，具体取决于任务延迟成本、设备利用情况及供电合同安排。

因此，AI 负荷的调节更需要依靠任务调度与能源配置协同完成。Google 于 2026 年 3 月披露，已与美国多家公用事业公司签订合计 1GW 的数据中心需求响应安排，通过限制或转移部分机器学习任务调节用电。我们认为，未来 AI 企业将同时追求服务连续性和用电灵活性，通过长期购电协议、储能、备用电源及负荷管理，降低电力约束对算力投运和持续运营的影响。

图5：AI 用电需求难以通过电价去调节，更加刚性



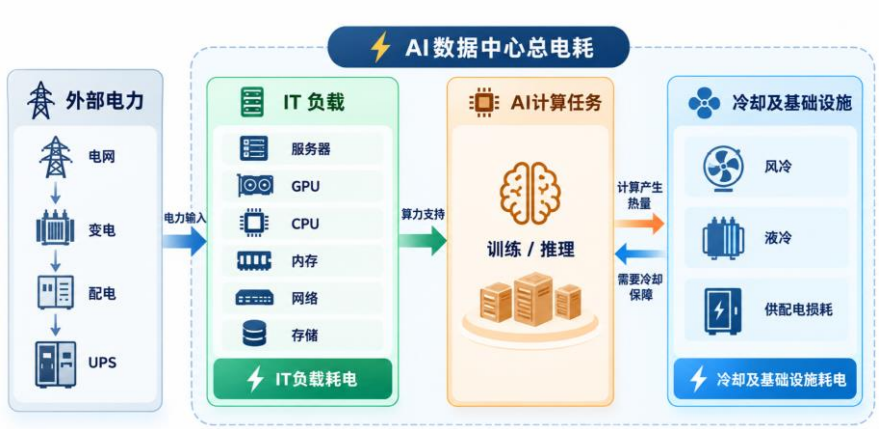
资料来源：国泰海通证券研究

3. AI 耗电拆解：从模型计算到数据中心电力消耗

我们认为，理解 AI 能源需求，需要建立从计算任务、硬件运行到数据中心总用电的完整传导链条。训练用于形成和更新模型能力，推理将模型能力转化为面向用户的服务，两者均依赖芯片、内存、网络和服务器协同运行，并进一步带来制冷及供配电需求。随着 AI 应用扩大，推理正在增加持续性的计算需求，模型迭代也将继续支撑训练投入。最终电耗取决于计算任务规模与单位任务能耗的共同变化，需要同时考察需求增长、硬件能效、软件效率及数据中心基础设施效率。

仅供内部参考，请勿外传

图6：AI 数据中心总电耗



资料来源：国泰海通证券研究

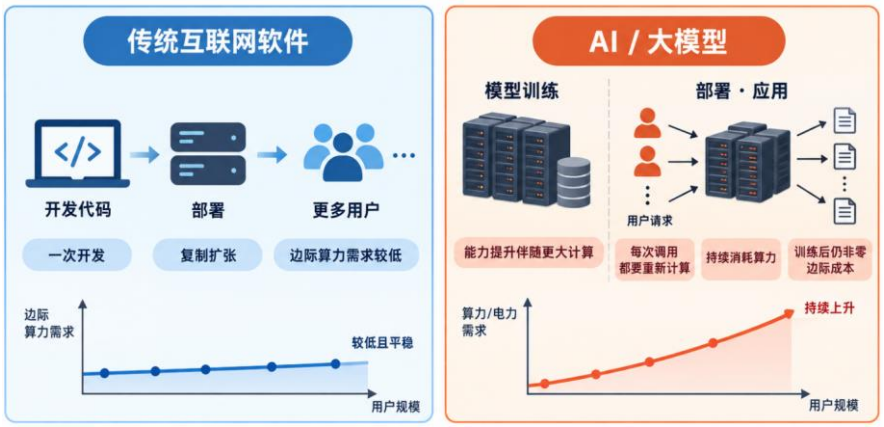
3.1. 从算法到算力：AI 运行依赖真实的物理载体

我们认为，AI 虽然以模型、算法和软件的形态呈现，但其运行并不是发生在虚拟世界，而是必须依赖真实的计算资源完成；无论是模型训练还是用户发起的一次推理请求，本质上都是将输入数据按照模型既定的计算规则进行大量数学运算，并最终生成输出结果。因此，AI 能力的实现首先需要“计算”，而计算是由芯片、服务器以及数据中心等物理设备承载。

过去市场对于 AI 的讨论更多集中在模型参数、算法能力和 GPU 数量，但三者之间实际上存在一条清晰的传导链条：模型决定需要完成什么样的计算任务，计算任务决定所需要的算力规模，而算力最终必须由真实的硬件设备完成。GPU 及其他 AI 加速芯片提供计算能力，服务器将大量芯片组织在一起，网络和存储系统负责不同设备之间的数据交换，而数据中心则进一步为这些设备提供可以长期、稳定运行的物理空间。这意味着“算力”并不是一个抽象概念，而是大量真实硬件协同运行之后形成的一种生产能力。

从这个角度看，AI 服务的规模化扩张具有持续消耗计算资源的特征。传统互联网服务同样需要服务器、网络和存储支持，而大模型在处理用户请求时，还需要执行相应的模型计算，使推理成本成为服务交付中的重要组成部分。不过，调用增长并不必然带来同比例的算力和电力增长：模型压缩、缓存、硬件升级及软件优化，都可能降低单次任务的资源消耗。AI 总电耗的变化，取决于服务需求扩张与单位任务能耗下降之间的相对速度。

图7：AI 用电需求与传统互联网服务业不同



资料来源：国泰海通证券研究

3.2. 训练与推理：AI 算力的两种基本形态

仅供内部参考，请勿外传

我们认为，理解 AI 算力首先需要区分“训练”与“推理”两种完全不同的计算过程。市场早期对于 AI 算力的关注更多集中在模型训练，大模型参数规模不断扩张带动 GPU 集群规模快速提升；但随着模型逐步从研发阶段进入大规模商业化应用，AI 算力的需求正在由单一的“训练模型”，进一步扩展至训练完成后的持续推理。两者虽然都依赖芯片完成计算，但在计算目的、运行方式以及对基础设施的要求上存在明显差异。

训练的本质，是通过大规模计算形成模型能力。模型在训练阶段需要对海量数据进行反复计算，并根据计算结果持续调整内部参数，最终形成能够理解和生成内容的模型。由于这一过程需要在大量 GPU 之间持续进行数据交换和并行计算，因此训练天然追求更大的集群规模和更高的算力密度。其计算任务相对集中，一次大型模型训练往往持续数周甚至数月，在训练完成之前，大量计算资源需要持续投入。也正因此，我们看到过去几年 AI 基础设施扩张首先表现为大型 GPU 集群和超大规模训练数据中心快速建设。

而推理则是将已经形成的模型能力真正转化为 AI 服务。模型完成训练以后，并不会像传统软件一样被无限复制后近乎零成本运行；当用户提出问题、生成图片、调用 Agent 或者企业通过 API 使用模型时，模型都需要重新执行一次计算过程，根据用户输入实时生成结果。也就是说，训练完成的是“模型”，而推理才是真正持续“生产 AI 服务”的过程。

训练与推理在任务组织和基础设施要求上存在差异。单次大型训练通常具有相对集中的运行周期，可以提前规划资源，并对集群互联和并行效率提出较高要求；推理则由持续发生的用户调用驱动，其中实时业务需要兼顾响应时延、峰值并发和服务可用性，离线任务则具有更大的调度空间。因此，分析 AI 负荷时，需要进一步区分不同任务的实时性、运行强度和可调节能力。

我们认为，AI 商业化将形成训练与推理共同扩张的需求结构。对于单个模型，训练形成能力，推理完成交付；对于整个行业，模型更新、后训练和新模型研发仍会持续发生，应用普及也将带来更多推理任务。因此，长期算力及能源需求应同时考虑模型研发投入、应用调用规模和技术效率变化，避免将训练视为一次性结束的投入，或将推理增长直接等同于电耗增长。

图8：AI 的训练与推理均需要持续用电



资料来源：国泰海通证券研究

3.3. 从算力到数据中心：AI 计算最终落到实际的能源电力消耗

我们认为，AI 虽然以软件和算法的形态呈现，但其生产过程本质上仍是一个高度物理化的能源转换过程：电力通过数据中心基础设施输入，驱动芯片完成计算，最终形成可被模型调用的算力。过去市场研究 AI 更多聚焦于模型参数、GPU 数量及理论算力，对于电耗的视角也主要在简单的“GPU 数量

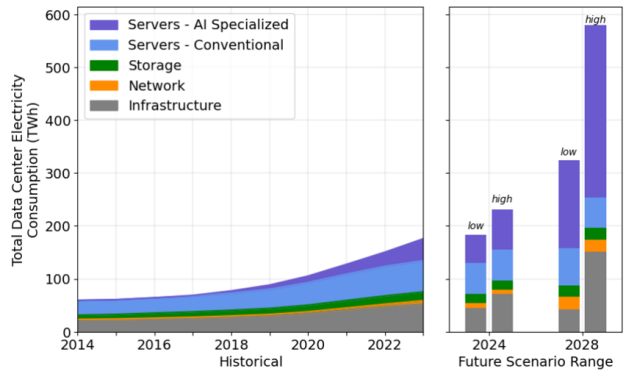
仅供内部参考，请勿外传

×单卡功率”；但随着 AI 从模型训练逐步进入大规模商业化应用阶段，仅从芯片维度已经难以解释 AI 真实的资源消耗。每一次模型训练和推理请求，背后都需要真实的服务器、网络、存储、供配电和冷却设施共同运行，**从能源视角研究 AI 需求需要建立整体化的研究框架。**

AI 算力需求最终需要通过真实的服务器和数据中心实现，AI 服务器的快速扩张已经开始重塑数据中心的电力消耗结构。根据美国劳伦斯伯克利国家实验室（LBNL）统计，美国服务器年用电量由 2014 年的约 30TWh 提升至 2023 年的接近 100TWh，其中 GPU 加速 AI 服务器用电量从 2017 年的不足 2TWh 快速提升至 2023 年的超过 40TWh，已经成为过去几年服务器用电增长最重要的来源之一。同期，美国数据中心整体用电量由约 60TWh 提升至 176TWh，占美国总用电量的比例达到 4.4%。我们认为，这一变化说明 AI 对于电力系统的影响并非简单来自数据中心数量增加，而是数据中心内部计算设备正在发生由传统 CPU 服务器向高功率 AI 加速服务器的结构性迁移。

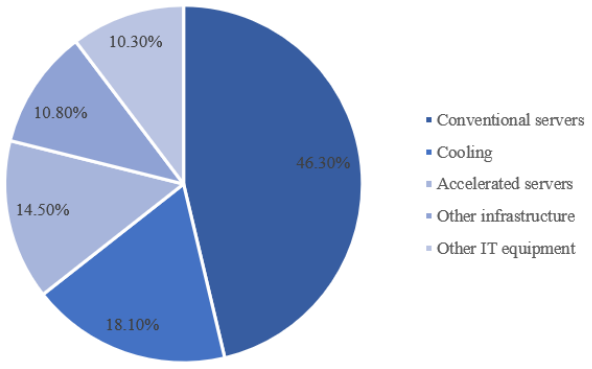
但 AI 数据中心的电力消耗并不等同于 GPU 耗电，服务器只是整个数据中心电力系统的核心环节之一。根据 IEA《Energy and AI》，2024 年现代数据中心中服务器平均占总用电量约 60%，存储系统约占 5%，网络设备最高约占 5%；除此之外，制冷系统占比根据数据中心类型和效率差异较大，在高效 Hyperscale 数据中心中约为 7%，而在效率较低的企业级数据中心中可以超过 30%。因此，从电网侧观察 AI 负荷，不能简单使用“GPU 数量×单卡功率”进行测算，而需要进一步考虑服务器、网络、存储、制冷以及供配电等完整数据中心系统。

图9：美国 AI 服务器快速增长成为数据中心新增用电的重要来源



资料来源：LBNL

图10：2024 年服务器占现代数据中心用电约 60%（传统+加速），制冷等基础设施仍占重要比例



资料来源：IEA，国泰海通证券研究

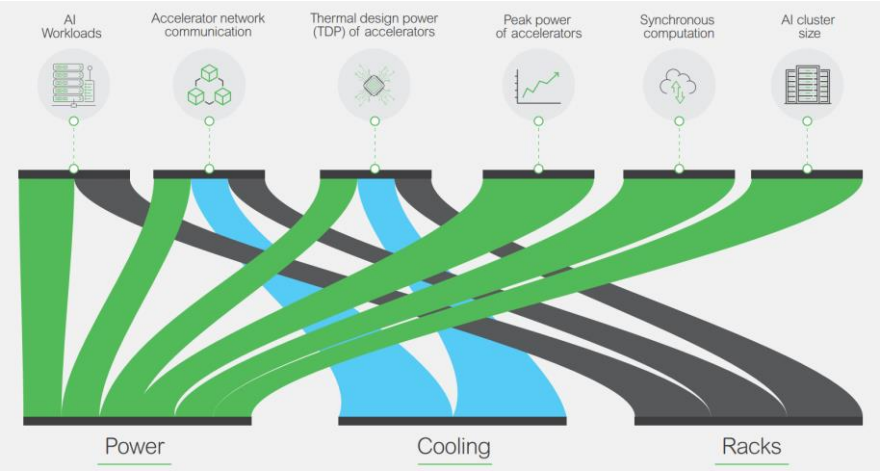
并且我们认为 AI 正在显著提高单位物理空间内的算力和功率密度，对数据中心供配电与散热系统提出完全不同于传统计算的要求。Schneider Electric 指出，当前 AI-ready 数据中心已经面临超过 100kW/机柜的功率密度，部分大型 AI 集群甚至达到数百 kW/机柜，传统数据中心的供电和风冷架构逐渐难以满足要求。以 NVIDIA GB200 NVL72 为例，单个机柜集成 72 颗 Blackwell GPU 和 36 颗 Grace CPU，官方公布整柜功率约 120kW，并采用以液冷为主的散热架构。也就是说，芯片性能和算力密度的提升正在直接向机柜、电力系统和制冷系统传导，AI 基础设施扩张已不再是简单“增加服务器”，而是数据中心物理架构的一次整体升级。

从数据中心整体效率来看，行业通常使用 PUE（Power Usage Effectiveness）衡量总用电与 IT 设备用电之间的关系。LBNL 数据显示，随着大型 Hyperscale 和 Colocation 数据中心占比提升以及基础设施效率改善，美国数

仅供内部参考，请勿外传

据中心基础设施用电占总用电比例已由 2014 年的约 40%下降至 2023 年的约 30%；但与此同时，由于 AI 服务器规模快速扩张，基础设施用电的绝对规模仍将继续上升。IEA 预计，在基准情景下全球数据中心用电量将由 2024 年的约 415TWh 提升至 2030 年的约 945TWh，其中 AI 驱动的加速服务器用电年均增速约 30%，贡献 2024—2030 年全球数据中心新增用电的接近一半；制冷、供配电等基础设施仍将贡献约 20%的新增用电。**我们认为随着 AI 算力规模和功率密度不断提高，电力与数据中心基础设施也由过去算力扩张的后台配套，逐渐成为决定算力能否真正落地的重要物理约束。**

图11：AI 算力扩张正在从芯片端向数据中心完整物理基础设施传导



资料来源：Schneider

4. AI 成本竞争，短期看硬件，实际看系统利用率，中长期看能源

我们认为，AI 算力成本需要从硬件投入、系统产出与能源条件三个维度共同考察。当前 GPU 及服务器仍占典型 AI 数据中心年度总成本的主要部分，芯片性能和采购价格决定成本基础；网络互联、软件优化及设备利用率，则决定这些投入能够形成多少实际计算产出。电力同时具有运营成本和供给约束两重属性：电价影响持续运行成本，供电容量与接入时间影响项目能否投运。随着模型能力、硬件获取和系统效率逐步进入可比区间，稳定、低成本且能够持续扩张的能源供给，有望成为区域 AI 竞争力的重要来源。

4.1. AI 数据中心成本结构：硬件是核心，运营成本中电力能源是大头

我们认为，考量 AI 数据中心的成本，首先需要将研究口径从“GPU 价格”扩展至整个数据中心的全生命周期成本（TCO）。AI 算力最终依赖 GPU、服务器、网络、电力及制冷等完整基础设施共同实现，因此一套 AI 算力的真实成本并不等于芯片采购成本，而是由前期资本开支（CAPEX）与运行期持续发生的运营成本（OPEX）共同构成。我们的研究表明，当前 GPU 等计算硬件仍是成本主体，但随着数据中心规模向百 MW 甚至 GW 级扩张，供电、制冷、土地和能源等过去处于后台的成本，也正在成为不可忽视的组成部分。

我们总结了 AI 数据中心年化成本的公式：
AI 年度 TCO = 年化计算硬件 CAPEX + 年化数据中心基础设施 CAPEX + 电力成本 + 网络、运维等 OPEX。

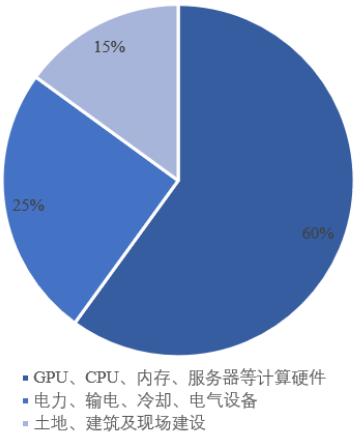
从资本开支角度看，根据 McKinsey 测算，为满足全球 AI 算力需求，到 2030 年仅 AI 相关数据中心就需要累计约 5.2 万亿美元资本投入；其中 GPU、

仅供内部参考，请勿外传

CPU、内存、服务器及机柜等计算硬件占总投资的约 60%（约 3.1 万亿美元）；电源、输电、冷却及电气设备占总投资的 25%（约 1.3 万亿美元），另外约 0.8 万亿美元、占 15%用于土地、建筑材料及数据中心建设。这已经非常清晰地表明 AI 数据中心中，计算硬件之外的基础设施投资仍然占到总资本开支约 40%，单纯使用“GPU 数量×GPU 价格”衡量 AI 算力成本，会明显低估真实投入。

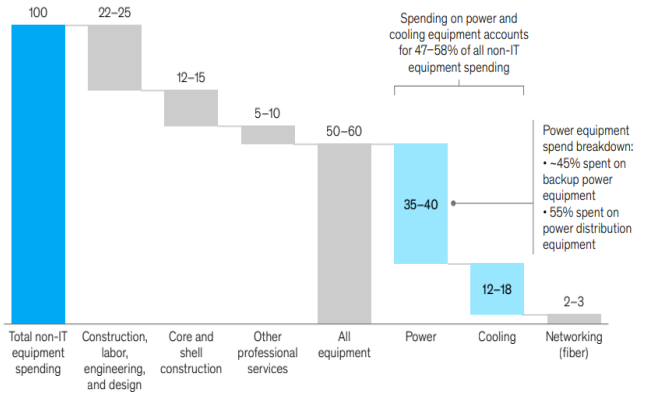
McKinsey 预计，到 2030 年全球数据中心累计资本开支将达到约 6.7 万亿美元，其中 AI 工作负载对应约 5.2 万亿美元；仅 AI 数据中心所需的供电与制冷设备累计投入就将达到约 7200 亿美元，其中供电设备约 5000 亿美元、制冷设备约 2200 亿美元。从非 IT 设备资本开支内部来看，供电设备占比约 35%—40%，制冷设备占比约 12%—18%。我们认为，这意味着 GPU 性能和功率密度的提升并不是孤立的硬件升级，而是会同步向上游变压器、配电、UPS、液冷等设备传导，AI 基础设施建设已经由“买 GPU”演变为完整的数据中心系统投资。

图12：AI 数据中心中，计算硬件之外的基础设施投资占到总资本开支约 40%



资料来源：McKinsey，国泰海通证券研究

图13：McKinsey 预计 AI 数据中心非 IT 的资本开支中电源及制冷占其中 47-58%，是最大头



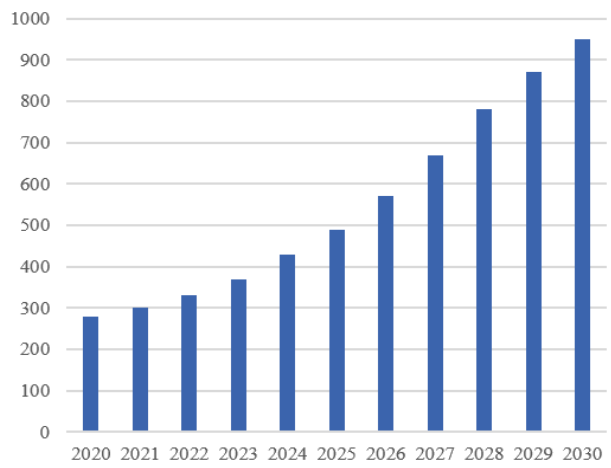
资料来源：McKinsey

我们认为除开资本开支之外，AI 数据中心电力等运营成本研究的重要性也越发凸显。IEA 数据显示，2024 年全球数据中心用电量已经达到约 415TWh，过去五年年均增长约 12%；基准情景下预计 2030 年进一步提升至约 945TWh，2024—2030 年复合增速约 15%。其中主要由 AI 驱动的加速服务器用电预计年均增长约 30%，并贡献未来全球数据中心新增用电量的接近一半。随着 AI 由集中式训练走向大规模推理，算力资产的运行时间和调用规模持续提升，电力也将由数据中心建设阶段的配套条件进一步转化为 AI 长期运营中的持续生产成本。

如果剔除 IT 硬件、单独观察一座 AI 数据中心的物理设施成本，电力已经成为最核心的生命周期成本。McKinsey 以一座 100MW AI Colocation 数据中心测算，中国数据中心平准化设施成本中约 55%来自 Utilities（主要为电力），接近 30%来自供配电和制冷设备，其余为建筑成本；不同地区由于电价和建设成本差异，AI 数据中心平准化设施成本相差较大。

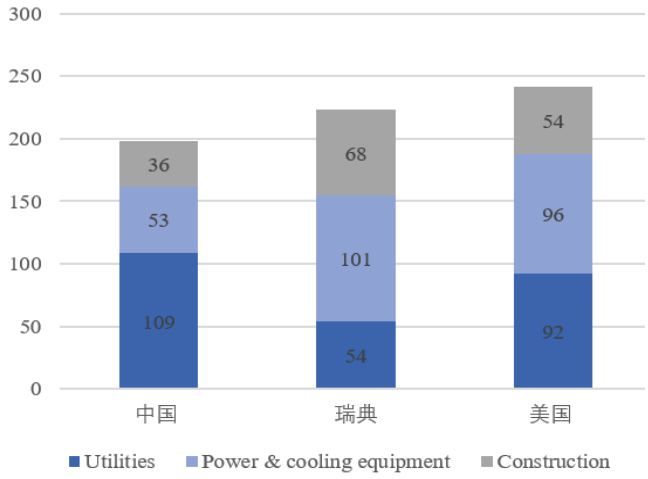
仅供内部参考，请勿外传

图14：IEA 预计全球数据中心电力消耗 2030 年相对于 2025 年翻倍（TWh）



资料来源：IEA，国泰海通证券研究

图15：建设和运营人工智能托管数据中心的最大成本是公用事业价格和资本支出，中国 55%成本与公用事业相关（$/MWh of facility energy）



资料来源：McKinsey，国泰海通证券研究

4.2. 短期看硬件：当前 AI 算力成本仍由芯片与服务器主导

我们认为，当前 AI 算力成本竞争的第一变量仍然是硬件，从完整 TCO 结构看，GPU、服务器及相关计算设备仍占据绝对主体。

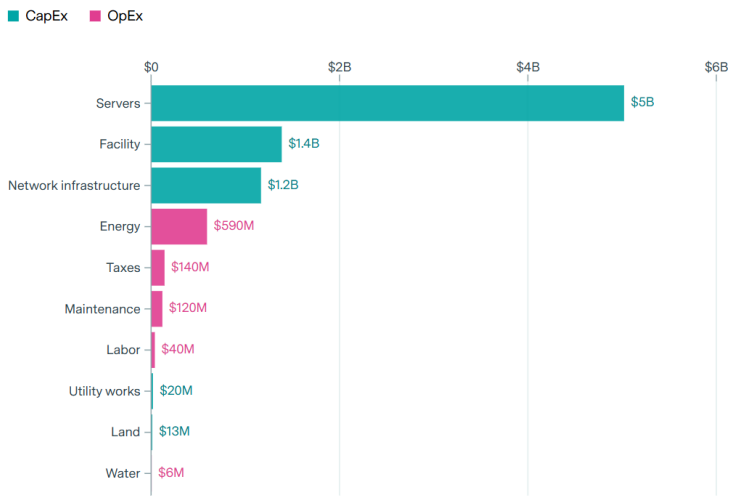
我们参考 Epoch AI 对一座 1GW IT 负荷、采用 NVIDIA GB200 NVL72 的典型 AI 数据中心测算，其年度 TCO 约 85 亿美元，其中服务器年化成本约 50 亿美元，占比接近 60%；数据中心设施及网络基础设施合计约占 30%，而年度电力成本约 5.9 亿美元，占比仅约 7%。这意味着，在当前技术和成本结构下，芯片采购价格、单位芯片性能以及完成同等任务所需要的硬件数量，仍然比电价本身更能决定 AI 算力成本。

我们认为短期硬件成本占主导也意味着，AI 基础设施不能简单以“单卡价格”衡量经济性，而需要比较完成相同任务所需要投入的整体硬件资本。如果一种芯片采购价格较低，但为了实现相同模型性能或吞吐量需要部署更多 GPU，则服务器、内存、网络互联以及配套供电和制冷设施均会同步增加，低单价优势可能迅速被更大的系统规模抵消；反之，更高价格的芯片如果能够显著提升单位芯片有效产出，也可能通过减少服务器数量和集群规模降低整体 TCO。

所以从短期看，决定 AI 算力成本的核心仍在硬件，我们以 Epoch AI 上述 1GW 数据中心模型为基础，其基准电价约 8.34 美分/kWh，年度电力成本约 5.9 亿美元。在其他条件完全不变的静态测算下，即便电价下降一半，年度 TCO 降幅仅约 3%—4%；即便将电价直接降至零，整体 TCO 理论降幅也仅约 7%。因此，如果不同硬件方案在单位有效算力成本上存在 10%甚至更大的差异，单纯依靠较低的电价通常难以完全弥补。

仅供内部参考，请勿外传

图16：AI 数据中心年度 TCO 仍以服务器成本为主



资料来源：Epoch AI

4.3. 实际看效率：理论算力并不等于真实产出

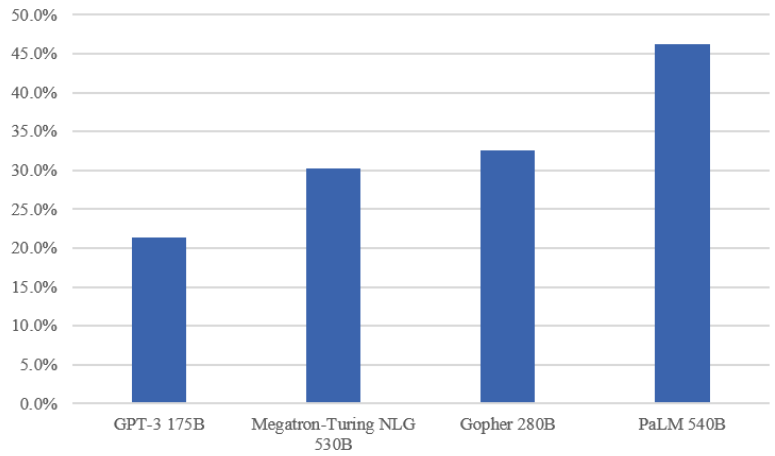
我们认为 AI 竞争短期看起来是在比谁拥有更多、更先进的 GPU，但在实际运行过程中系统效率才是决定硬件最终能够转化为多少有效计算产出的核心，比拼谁能够以更高效率将这些硬件转化为真实的训练结果、Token 和 AI 服务。

GPU 标称算力反映的是理想条件下的理论性能上限，而实际模型运行过程中还受到显存带宽、内存访问、芯片互联、网络通信、数据同步、并行策略以及软件框架等多重因素约束。模型越大、使用 GPU 等加速芯片数量越多，整个系统的协同效率对实际产出的影响越显著。根据 Chowdhery 等（2023）发表的 PaLM 论文，GPT-3、Megatron-Turing NLG、Gopher 及 PaLM 四个大型模型的 Model FLOPs Utilization（模型浮点运算利用率，MFU）分别为 21.3%、30.2%、32.5%和 46.2%，四个模型的中位数约为 31.35%。其中，Google 训练 5400 亿参数 PaLM 模型时，采用高度优化的大规模 TPU 集群，MFU 达到 46.2%。这意味着，同样拥有 100 单位理论算力，不同系统最终真正用于模型计算的有效算力可能存在显著差距。理论算力决定硬件能够达到的上限，而系统效率决定这一上限究竟能够释放多少。

所以我们认为，单位有效算力成本应该归纳为：AI 数据中心 TCO ÷ 实际有效计算产出。

仅供内部参考，请勿外传

图17：大型模型实际有效算力显著低于理论峰值



资料来源：Chowdhery 等（2023），《PaLM: Scaling Language Modeling with Pathways》，国泰海通证券研究

这意味着硬件投入只是成本的一侧，系统效率则直接影响产出的另一侧。相同 GPU 通过更高带宽的芯片互联、更成熟的并行策略、更高效的软件框架，以及量化、Batching、Speculative Decoding 等优化手段，可以在几乎不增加硬件投入的情况下提高实际吞吐；反过来，即使芯片理论性能领先，如果通信、内存或者软件效率不足，其实际单位成本仍可能明显高于理论参数所显示的水平。

利用率又进一步决定这些高额固定资产能够被多少实际任务所摊薄。GPU、服务器、网络和机房一旦建成，其折旧和资本占用已经发生，无论设备是否持续工作，固定成本都不会消失。所以类比传统制造业，更大的产量必然带来单位成本的极大摊薄，这也会平抑掉很多总投资方面的差异。

另外，训练和推理在这一问题上又存在明显差异。大型训练任务可以提前规划，并在数周甚至数月内连续占用集群，理论上更容易形成高负载；而推理需求直接由终端用户调用驱动，需要兼顾日内流量波动、峰值并发和响应时延。为了保证服务质量，推理数据中心往往需要预留一定冗余算力，由此形成“按峰值配置、按平均需求使用”的结构性矛盾。随着 AI 由训练逐渐进入大规模应用阶段，调度能力、负载匹配和利用率将越来越直接地决定 AI 基础设施的真实单位经济性。

图18：NVIDIA 用完全相同的 H200 GPU + Llama 3.1 70B 做实验，相同 H200 GPU 条件下，引入 NVSwitch，特定测试配置吞吐最高提升约 50%

批量大小	吞吐量 tok/s/GPU		NVSwitch 优势
	点对点	NVSwitch	
1	25	26	1.0x
2	44	47	1.1倍
4	66	76	1.2倍
8	87	110	1.3倍
16	103	142	1.4倍
32	112	168	1.5倍

资料来源：Nvidia

4.4. 能源影响扩张能力与区域竞争力，硬件成本被持续摊薄后，能源的重要性将逐步上升

我们认为，从短期角度看能源的供应能力决定了区域内 AI 需求的扩张天花

仅供内部参考，请勿外传

板；而当硬件采购条件、模型能力和系统效率趋于可比时，电价、PUE 及供电条件的差异，更可能成为区域成本竞争的决定因素，能源在长期角度的重要性更不可忽视。

我们观察到 2024 年蔓延的“美国缺电”问题，电力供给已经影响当前算力投运，其经济价值可能显著超过电费在成本中的占比。对于高资本投入的数据中心，即使电费只占年度总成本的一小部分，接电延迟或供电容量不足，仍可能使服务器和机房无法形成预期产出。由此，评价能源条件需要同时考察“电价多低”和“何时能够获得足额、稳定的电力”。能够提前落实容量、缩短接入周期并保障持续运行的地区，有助于减少资产等待时间，提高资本投入转化为有效算力的速度。

过去几年 AI 芯片单位美元性能持续快速提高。Epoch AI 统计显示，自 2023 年以来市场实际采购 AI 芯片的单位美元性能平均每年提升约 49%，约每 1.7 年翻倍；2025 年采购芯片的平均单位美元性能较 2024 年进一步提高约 91%。这意味着，同样一美元硬件资本能够购买到的计算能力仍在快速增长，硬件成本天然具有明显的技术摊薄属性。

与此同时，算法、模型架构和软件优化还在进一步提高单位硬件能够完成的实际计算任务。硬件价格下降与有效产出提高共同作用，使单位计算所需要承担的资本成本持续下降。但电力并不会因为 GPU 采购成本下降而同步归零：当芯片性能提高之后，数据中心往往不是简单维持原有算力规模，而是进一步部署更多设备、运行更大的模型并处理更多推理请求。IEA 预计全球数据中心用电量将由 2024 年的约 415TWh 增加至 2030 年的约 945TWh，其中 AI 加速服务器贡献接近一半新增用电。这意味着，单位计算能效提升并没有阻止整个 AI 产业总能源需求快速增长。

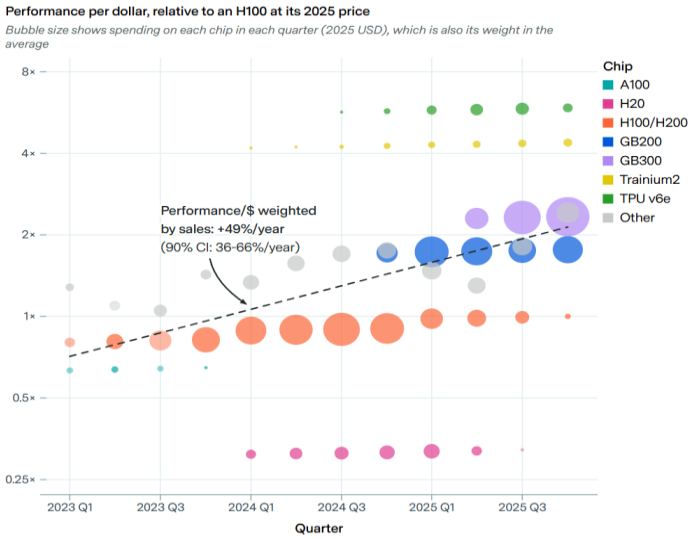
推理时代将进一步强化这一特征。模型训练具有一定阶段性，一次大型训练完成以后，模型能力已经形成；但推理则伴随用户数量和 AI 应用规模持续发生。用户每提出一次问题、生成一段视频、调用一次 Agent 或者企业通过 API 完成一次任务，模型都需要重新执行计算。AI 越成功、用户越多、Token 生成规模越大，对算力和能源的需求反而越高。

因此，我们认为电力优势具有明显的“后发性”。在硬件差距较大的阶段，更先进的 GPU 和更高的系统效率可以轻易覆盖不同地区几个百分点的电力成本差异；但当芯片性能、采购价格和系统效率逐渐进入可比区间以后，电价、PUE 以及供电可靠性的影响将迅速放大。对于 7×24 小时持续运行的大规模推理基础设施而言，即使单位计算耗电持续下降，长期、连续运行所形成的累计能源成本仍将越来越重要。

更重要的是，能源优势相较硬件和软件更难跨区域复制。领先 AI 芯片可以通过资本投入和贸易进行配置，算法与软件可以快速复制，但一个地区的电价水平、电源结构、电网容量和供电可靠性由长期形成的能源系统决定，很难在几年时间内快速改变。**因此，当全球 AI 硬件和系统效率差距逐步收敛之后，低成本、稳定且能够大规模扩张的电力资源，有可能从当前的辅助变量逐步转变为长期 AI 成本竞争中更难复制的核心禀赋。**

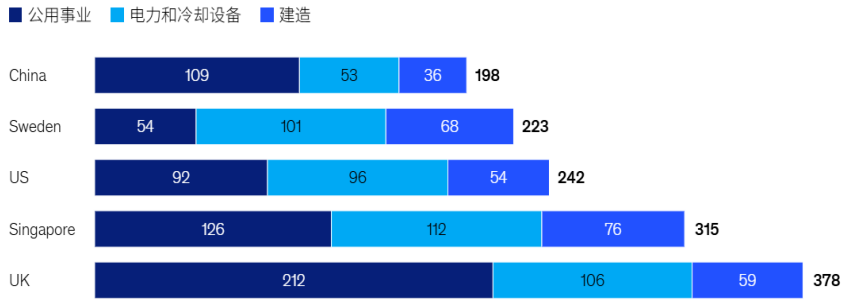
仅供内部参考，请勿外传

图19：Epoch AI 最新统计显示，2023 年以来实际采购 AI 芯片的 Performance per Dollar 平均每年提高约 49%



资料来源：Epoch AI

图20：AI 托管数据中心能源的平准化成本，样本中国设施成本约为英国的 52%，公用事业成本解释约 57%的差距 （美元/兆瓦时设施能源）



资料来源：McKinsey

5. 从数据中心成本到 AI 商业化

我们认为，随着 AI 应用加快普及，产业研究需要进一步从算力和数据中心投入，延伸至服务产出与商业回报。推理将模型能力持续转化为面向用户的服务，Token 由此成为连接使用规模、商业定价和底层成本的重要计量单位；在模型质量与服务要求可比的条件下，将推理业务对应的基础设施成本与实际 Token 产出相匹配，可以进一步识别硬件、系统效率、设备利用率及能源条件对单位服务成本的影响。由此，AI 竞争在持续追求模型能力突破的同时，也将更加重视商业化交付的经济性：先进算力决定能力边界，系统效率影响成本与产出，稳定且低成本的能源供给则支撑持续运营与规模扩张。能否将算力投入转化为具有成本竞争力的服务，并形成可持续的商业回报，将成为衡量 AI 企业长期竞争力的重要标准。

仅供内部参考，请勿外传

图21：AI 竞争重心未来可能由算力规模逐步向 Token 经济性与能源迁移



资料来源：国泰海通证券研究

5.1. AI 商业模式推进，token 成为经济性计量单位

我们认为，随着 AI 服务规模化交付，Token 正在成为连接模型使用量、基础设施投入与商业定价的重要计量单位。训练和模型更新形成能力，推理将这些能力转化为面向企业与个人用户的服务；持续发生的调用，则对应实际计算消耗及相应成本。因此，研究 AI 经济性需要进一步从算力投入延伸至单位服务产出，考察既定模型质量和服务要求下，每单位 Token 需要承担多少基础设施成本，以及这些产出能够形成多少商业收入。

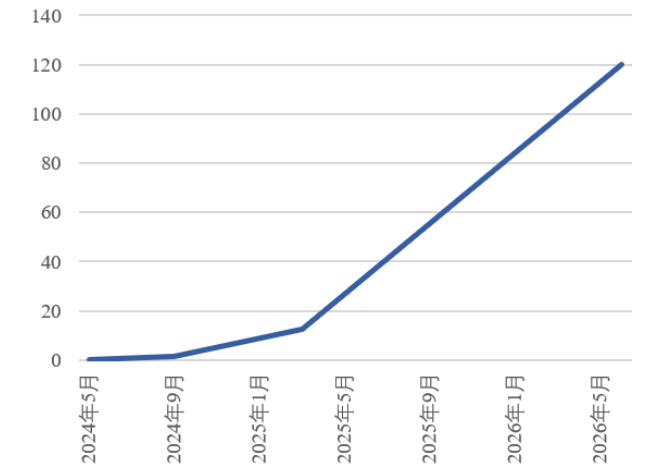
训练阶段主要表现为 GPU、服务器及数据中心的大规模前置投入，而模型训练完成后，只有通过持续推理才能将已经形成的模型能力转化为问答、搜索、代码生成、Agent 等实际服务。对于 B 端而言，模型通过 API 嵌入企业软件和业务流程；对于 C 端而言，则通过 Chatbot、搜索、办公、内容生成等产品直接触达用户。与传统软件一次开发后可以接近零边际成本复制不同，AI 服务每发生一次调用，都需要重新消耗算力完成推理，模型商业化越成功，反而意味着后续需要生产的 AI 服务越多。

并且 AI 模型的成功与算力需求并不存在传统软件意义上的“开发完成后成本趋近于零”，而是可能形成明显的正反馈。模型能力越强，可进入的应用场景越多；用户和企业使用规模扩大，又进一步提高调用频率；长上下文、深度推理、多模态和 Agent 的发展，则继续增加单次任务所需要处理的 Token 数量。由此，AI 推理需求可以近似理解为用户数量 × 调用频率 × 单次调用规模的共同结果，模型商业化越成功，后续 Token 处理量和持续算力需求可能反而越大。

Google 披露，其旗下模型在产品和 API 中的月度处理量由 2024 年 5 月的约 9.7 万亿 Token 增长至 2026 年 5 月超过 3.2 千万亿 Token，两年增长超过 300 倍；国内豆包大模型日均 Token 调用量则由 2024 年发布初期约 1200 亿增长至最新约 120 万亿，增幅约 1000 倍。模型能力提升不仅带来更多用户，同时推动调用频率提高、上下文拉长以及 Agent、深度推理等复杂任务增加，使 AI 需求逐步形成“用户数量×调用频率×单次任务计算量”共同驱动的增长模式。

仅供内部参考，请勿外传

图22：豆包日均 Token 调用量爆发增长（万亿）



资料来源：新华网，火山引擎 Agent 开发社区，国泰海通证券研究

图23：Google 月度 Token 处理量两年增长约 330 倍



资料来源：GOOGLE

与此同时，Token 正在将这种持续性的推理服务进一步转化为可以计量和定价的商业产出。目前主流大模型 API 已经普遍按照每百万 Input/Output Token 进行收费。例如 OpenAI 当前 API 定价即以每 100 万 Token 作为核心计费单位，并分别对输入、缓存输入和输出 Token 进行定价。 C 端产品虽然更多通过会员订阅、广告和增值服务实现收入，但其底层同样对应持续发生的 Token 生成和推理成本。因此，Token 开始同时具备 AI 服务产出计量和 B 端商业定价两重属性。

我们认为，这使得 AI 产业逐步形成一套完整的商业循环：前期资本投入形成模型与算力能力，推理将模型能力交付给 B 端和 C 端用户，持续增长的 Token 调用进一步转化为 API 收入、订阅收入及其他商业回报，并反过来支撑下一轮模型训练、芯片采购和数据中心投资。

图24：主流模型 API 按 Token 计价

模型名称	厂商	类型	输入价格 $/1M tokens	输出价格 $/1M tokens	上下文长度
GPT 6.1 Sol	OpenAI	T I → T	$2.00	$10.0	1.1M
GPT-6 Astra	OpenAI	T I → T	$10.0	$50.0	1.1M
GPT-6 Luna	OpenAI	T I → T	$0.10	$0.50	1.1M
GPT-5.6 Sol	OpenAI	T I → T	$5.00	$30.0	1.1M
GPT-5.5 Pro	OpenAI	T I → T	$30.0	$180.0	1.0M
Claude Fable 5.1	Anthropic	T I → T	$10.0	$50.0	1.0M
Claude Opus 5.5	Anthropic	T I → T	$4.00	$20.0	1.0M
Claude Haiku 4.5	Anthropic	T I → T	$1.00	$5.00	200K
Gemini 4 Argon	Google	T I V A → T	$2.00	$10.0	1.0M
Gemini 3.8 Flash	Google	T I V A → T	$0.75	$3.75	1.0M
Gemini 3.5 Flash Lite	Google	T I V A → T	$0.30	$2.50	1.0M
Gemini 3.1 Pro	Google	T I V A → T	$2.00	$12.0	1.0M
Gemini 3.1 Flash Lite	Google	T I V A → T	$0.25	$1.50	1.0M

资料来源：LLM API Prices，国泰海通证券研究

5.2. 从数据中心 TCO 到 Token Cost：AI 成本竞争开始具备统一比较口径

我们认为，Token Cost 是连接 AI 物理基础设施与商业产出的核心桥梁。当

仅供内部参考，请勿外传

硬件、效率、利用率和能源成本最终统一映射到单位 Token 之后，AI 产业的竞争指标也将逐步由“算力规模”进一步转向“单位 token 生产成本”的竞争。由此，AI 成本竞争已初步具备建立全生命周期经济性分析框架的条件。

从成本端看，Token Cost 的分子是完整的 AI 数据中心 TCO，而非单纯的 GPU 采购价格。Epoch AI 对 1GW AI 数据中心的测算显示，其年度 TCO 约 85 亿美元，其中服务器年化成本约 50 亿美元，占比接近 60%；数据中心设施、网络以及持续发生的电力和运维成本共同构成剩余部分。对于 AI 企业而言，无论 GPU 是否持续满负荷运行，服务器折旧、数据中心设施和网络资本成本均已经发生，因此需要将整个 AI 基础设施的全生命周期成本统一纳入计算。

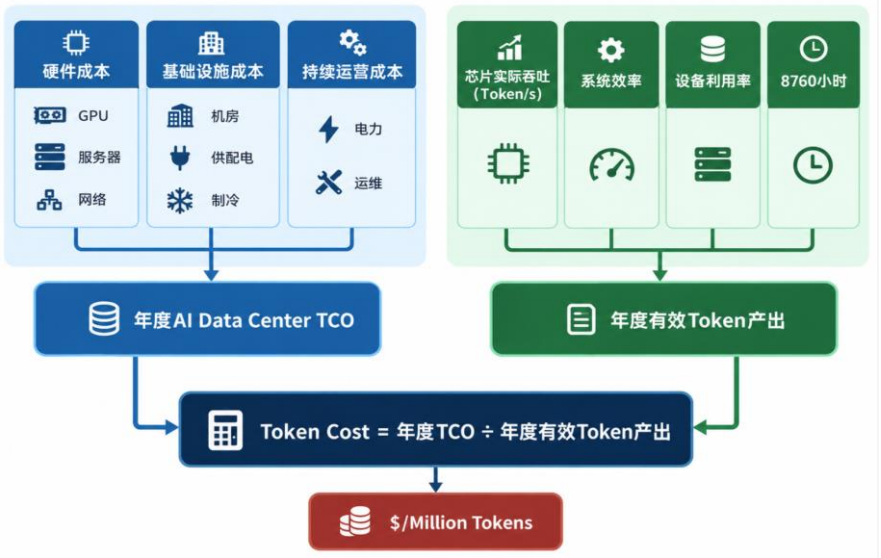
从产出端看，真正决定 Token 成本的并不是理论 FLOPS，而是数据中心一年能够实际生产多少有效 Token。同样数量的 GPU，在不同模型、互联架构、软件效率以及利用率下，其实际 Token 吞吐可能存在明显差异。因此可以将单位 Token 成本简化表示为：

Token Cost = 年度 AI 数据中心 TCO ÷ 年度有效 Token 产出

而年度有效 Token 产出取决于实际 Token 吞吐 × 有效运行时间 × 利用率。

我们认为，过去不同芯片、不同数据中心和不同地区之间难以直接比较，是因为不同区域和国别在 GPU 价格、算力性能、电价及 PUE 等方面存在较大差异；而一旦最终统一折算为美元/百万 Token（$/M Token），硬件、系统效率和能源禀赋便可以进入同一套经济比较框架。当然我们也认识到当前 Token 并非完全同质化产品。不同模型、输入与输出 Token、上下文长度、推理深度以及响应时延要求，对应的实际计算量并不相同。但通过 Token 成本的角度，至少可以给予我们一个较之前相对合理且公平的经济学计量。

图25：从 Data Center TCO 到 Token Cost



资料来源：国泰海通证券研究

5.3. 从算力规模到 Token 经济性：AI 竞争重心或将逐步由能力向成本迁移

我们认为，在前沿模型能力仍快速迭代、不同厂商模型性能尚未充分收敛的阶段，竞争首先仍然取决于谁能够获得并部署更大规模、更先进的算力。所以我们总结为短期看算力，决定谁能够站上模型能力前沿；中期看 Token 成

仅供内部参考，请勿外传

本，决定谁能够以更高效率实现商业化，获得经济性的正向循环；而能源当前影响算力能否及时落地，持续运营中影响成本与可靠性，在其他条件趋于可比时还可能形成中长期的区域竞争优势。

对于头部模型厂商而言，只要新增算力仍能够推动模型能力显著提升，争夺先进 GPU、扩大训练集群和建设超大规模数据中心的战略优先级，就可能高于短期单位 Token 成本优化。此前 xAI 的扩张进一步体现了这一特征。马斯克披露，截至 2026 年 9 月，Colossus 2 已经部署约 11 万张 GB200 和 44 万张 GB300，并计划再分别于近期、11 月和可能的 12 月底新增三批各 22 万张 GB300；若全部兑现，Colossus 2 年底将达到约 121 万张 GB200/GB300，其中 GB300 约 110 万张。我们认为这充分反映了当前前沿 AI 竞争仍然具有非常明显的“算力军备竞赛”特征：对于仍处于能力追赶和模型迭代阶段的厂商而言，大规模先进算力本身就是进入前沿模型竞争的重要门槛。

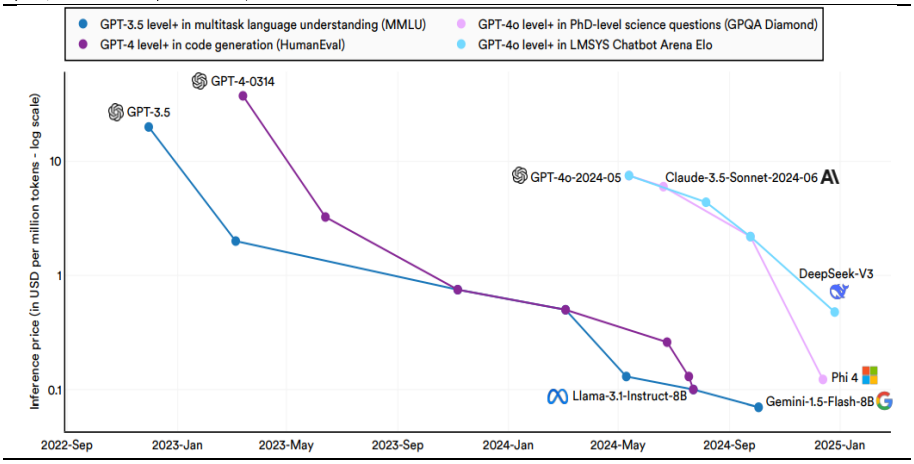
但我们认为，算力规模并不会永久成为唯一竞争维度。一方面，AI 硬件本身仍在快速进步。Epoch AI 测算显示，2023 年以来市场实际采购 AI 芯片的单位美元性能平均每年提高约 49%，约每 1.7 年翻倍；2025 年购买芯片的平均性价比较 2024 年提高约 91%。随着先进硬件持续扩散，单纯依靠更多 GPU 建立的相对优势可能逐步被技术迭代和资本投入所追赶。

另一方面，当模型能力逐步进入可比区间后，竞争焦点将从“模型能不能完成任务”进一步转向“以多低的成本完成同样的任务”。Stanford AI Index 数据显示，达到 GPT-3.5 相近 MMLU 能力水平的模型，其推理价格由 2022 年 11 月约 20 美元/百万 Token 下降至 2024 年 10 月约 0.07 美元/百万 Token，其间下降超过 280 倍；不同任务上的 LLM 推理价格年降幅达到 9—900 倍。快速下降的推理价格意味着，当模型能力差距收敛后，单位服务成本、价格和毛利率将越来越成为商业竞争的重要变量。

因此，我们认为，AI 竞争将随着应用深化，进一步增加对单位服务经济性的关注。在前沿能力仍存在明显差异的场景，算力投入和模型研发继续决定竞争门槛；在服务质量进入可比区间的场景，系统效率、设备利用率和单位交付成本的重要性将提高。电力供给贯穿算力建设与运营，低成本、可靠且能够持续扩张的能源条件，有望在其他投入趋于可比时形成稳定优势。从商业模式看，Token Price 与 Token Cost 仍需分别评价：较低报价可能来自效率提升，也可能来自补贴和利润让渡；可持续的竞争力，需要在满足用户需求的同时，使服务收入覆盖生产成本，并形成支持后续研发与基础设施更新的回报。

仅供内部参考，请勿外传

图26：同等能力模型的推理价格大幅下降，支撑“能力逐渐扩散后，中期竞争开始转向单位经济性”



资料来源：Stanford

仅供内部参考，请勿外传

6. 风险提示

AI 应用及推理需求增长低于预期。若大模型商业化进程、B 端/C 端渗透率及 Token 调用量增长低于预期，可能导致算力及数据中心需求增长不及预期。

AI 软硬件效率提升超预期。若芯片能效、模型压缩、推理优化及系统效率提升速度超预期，单位 Token 所需算力及电力消耗可能显著下降，使能源需求及成本竞争格局与本文判断存在偏差。

仅供内部参考，请勿外传

本公司具有中国证监会核准的证券投资咨询业务资格

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，保证报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业理解，本报告清晰准确地反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响，特此声明。

免责声明

本报告仅供国泰海通证券股份有限公司（以下简称“本公司”）授权客户使用。本公司不会因接收人收到本报告而视其为本公司的当然客户。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。

本报告的信息来源于已公开的资料，本公司对该等信息的准确性、完整性或可靠性不作任何保证。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌。过往表现不应作为日后的表现依据。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保持在最新状态。同时，本公司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。

本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。在任何情况下，本公司、本公司员工或者关联机构不承诺投资者一定获利，不与投资者分享投资收益，也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。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

本公司利用信息隔离墙控制内部一个或多个领域、部门或关联机构之间的信息流动。因此，投资者应注意，在法律许可的情况下，本公司及其所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。在法律许可的情况下，本公司的员工可能担任本报告所提到的公司的董事。

市场有风险，投资需谨慎。投资者不应将本报告作为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向专业人士咨询并谨慎决策。

本报告版权仅为本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制、发表或引用。如征得本公司同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“国泰海通证券研究”，且不得对本报告进行任何有悖原意的引用、删节和修改。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息或进而交易本报告中提及的证券。本报告不构成本公司向该机构之客户提供的投资建议，本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

评级说明

投资建议的比较标准

投资评级分为股票评级和行业评级。

以报告发布后的 12 个月内的市场表现为比较标准，报告发布日后的 12 个月内的公司股价（或行业指数）的涨跌幅相对同期的沪深 300 指数涨跌幅为基准。

	评级	说明
股票投资评级	增持	相对沪深 300 指数涨幅 10%以上
	谨慎增持	相对沪深 300 指数涨幅介于 5%～10%之间
	中性	相对沪深 300 指数涨幅介于-5%～5%
	减持	相对沪深 300 指数下跌 5%以上
	无评级	对于个股未来市场表现与基准指数相比无明确观点
行业投资评级	增持	明显强于沪深 300 指数
	中性	基本与沪深 300 指数持平
	减持	明显弱于沪深 300 指数

国泰海通证券研究所

地址 上海市黄浦区中山南路 888 号 邮编 200011 电话 （021）38676666