2025 年 03 月 21 日

通信行业研究

买入（维持评级）

行业深度研究

证券研究报告

通信组
分析师：张真桢（执业 S1130524060002）
zhangzhenzhen@gjzq.com.cn

DeepSeek 算力效率提升≠算力通缩，国产算力需求方兴未艾

投资逻辑：

DeepSeek 在知乎发布文章《DeepSeek-V3/R1 推理系统概览》，披露其 AI 大模型的理论成本利润率高达 545%，引发业内的热烈讨论。在本篇报告中，我们从以下三个角度：1）DeepSeek 的底层架构优化；2）DeepSeek 的利润率详细拆解；3）DeepSeek 引发的算力需求之争，回应市场关心的问题。此外，当前市场针对算力之争多定性分析，本篇报告也旨在提供较完整的定量分析框架以供参考。

DeepSeek 通过大规模专家并行与计算通信重叠提升算力效率：

大规模专家并行模式下，专家参数被存储在多个 GPU 中，集群处理并行请求能力得到增强，GPU 算力资源利用率也得到了提高。但在此模式下，通信耗时增加，因此 DeepSeek 还采用算通信重叠策略以缓解该问题。我们认为 AI 大模型具有规模效应：通过底层架构优化后，伴随批量大小的增加，计算与通信的时间边际下降，吞吐率得到提升。因此大规模集群能提高算力利用率。

参考 DeepSeek，我们认为 MaaS 厂商具有盈利潜力，率先实现规模效应的云厂商将脱颖而出：

我们对 DeepSeek 披露的 545%高利润率进行了拆解，以进一步分析利润率的影响因素。545%是成本利润率，对应 84.5%的收入利润率。将 GPU 租赁成本在总成本中的占比、付费率调至合理水平后，我们认为公司实际利润率或低于 84.5%。付费率对公司利润率影响较大，伴随付费率的提升，公司利润率有望持续攀升。若能将付费率提升至 40%+，则公司的利润率有望达 20%+。根据对 DeepSeek 的利润率分析，我们认为 MaaS 模式具有盈利潜力。拥有大规模集群、能形成高用户并发的公有云厂商有望形成规模效应。

针对算力之争，我们认为算力效率是新的 Scaling Law 方向，多模态与 AI Agent 将打开算力的成长空间：

我们就 DeepSeek 的模型参数量、数据规模、峰值倍数、单卡算力、单卡利用率等关键指标进行了详细的拆解，发现 DeepSeek 低算力的原因在于：1）低峰值倍数：未设置较大的算力冗余（峰值倍数仅 1.2），一定程度上牺牲了用户体验；2）超高的算力效率，具体体现在单次推理激活的模型参数量（单次推理仅激活 370 亿参数）、高单卡利用率（H800 单卡利用率高达 77%）。市场担心 DeepSeek 仅使用 1814 个 H800 就支持了约 2500 万 DAU 会证伪算力需求，但我们认为伴随峰值倍数的提高、数据规模的扩大，算力需求有望持续提升。高算力效率不等于算力通缩，“参数量*效率*数据规模”才是新的 scaling law 方向。从远期看，多品类 APP 接入 AI 大模型有望带来用户数的增长，多模态、AI Agent 有望带来单次请求调用 tokens 数量的增加，这都将带动算力需求的提升。因此我们持续看好算力链。

投资建议与估值

我们认为算力需求将持续强劲，建议持续关注算力链板块。能实现大集群、形成高用户并发的公有云厂商有望率先实现规模效应，跑通 MaaS 盈利模式。我们看好能提供高安全可靠的云服务、并具有辐射全国的 IDC 资源的运营商云；也看好具有丰富客户资源、集团内部生态赋能的互联网大厂云。深度参与算力产业链的国产芯片、交换机厂商也将持续受益。建议关注中国移动、中国联通、中兴通讯等。

风险提示

AI 落地不及预期、芯片供应不足、客户对公有云接受度不及预期、行业竞争加剧

内容目录

1、DeepSeek 通过大规模专家并行（EP）与计算通信重叠（DP），大幅提升算力效率 4
2、参考 DeepSeek，MaaS 厂商具有盈利潜力，率先实现规模效应的云厂商将脱颖而出 7
2.1. DeepSeek 口径下，545%高利润率的计算详解 7
2.2. 根据实际情况调整后，DeepSeek 利润率有所降低，我们认为付费率为关键影响因素 8
2.3. MaaS 模式具有盈利潜力，看好能形成规模效应的公有云厂商 11
3、算力需求之争：算力效率是新的 Scaling Law 方向，多模态与 AI Agent 将打开算力的成长空间 13
3.1. DeepSeek 的 1814 个 GPU：算力效率的提升、算力冗余和用户使用频次的不足 13
3.2. AI Chatbot 的算力需求量估算：我们预计约 60-70 万片 15
3.3. 多品类 APP 接入 AI 大模型、多模态、AI Agent 原生产品带动算力需求扩容 16
4、相关标的 18
5、风险提示 20

图表目录

图表 1： DeepSeek-V3/R1 采用了预填充-解码分离架构，有效提升算力效率 4
图表 2： DeepSeek-V3/R1 参考架构图 4
图表 3： 预填充阶段，将一批请求分成两个微批次 5
图表 4： 英伟达接入 DeepSeek-R1 后，实现 H200 和 B200 的吞吐能力提升 5
图表 5： DeepSeek 进行架构优化后，批量大小的增加将更有效地带动吞吐率的提升 6
图表 6： DeepSeek 口径下，其 GPU 租赁成本测算得到$87,072/天 7
图表 7： DeepSeek 口径下，其大模型收入测算得到约为$562,100/天 8
图表 8： 按利润率=利润/总收入计，DeepSeek 利润率约为 84.5% 8
图表 9： 大模型厂商的成本还包括运维、带宽、人力等其他成本 9
图表 10： 为计算简便，DeepSeek 在进行利润率估算时简化了三个因素 9
图表 11： DeepSeek 官网对不同时段、不同模型、不同情形的定价 10
图表 12： 假设 DeepSeek V3/R1 调用需求占比分别为 35%/65%，则得到各时段、各情形的均价 10
图表 13： 各时段、各情形下的 tokens 调用情况 10
图表 14： 不考虑付费率的情况下，DeepSeek 的利润率高达 66% 11
图表 15： DeepSeek 利润率敏感性分析 11
图表 16： DeepSeek 披露其过去的 24 小时内用于推理服务的 H800 数量约为 1814 个 13
图表 17： 我们推算得到 DeepSeek 对 H800 的单卡利用率高达 77% 13
图表 18： 影响算力需求的因素包括算力效率、算力冗余、用户使用强度 14
图表 19： 根据我们估算，当前单位 DAU 搜索请求次数仅 1.55 次 15

图表 20： AI Chatbot 应用的 AI 芯片需求量约为 66 万片 15
图表 21： AI Chatbot 应用的算力需求量敏感性分析 16
图表 22： Gemini2.0 Pro 可支持 2M 的上下文长度 16
图表 23： Manus 能自发执行任务 17
图表 24： 我们估算多模态、AI Agent 的发展将带动约 800 万的 H20 需求 17
图表 25： 相关标的估值（截至 2025 年 3 月 20 日收盘价） 19

1、DeepSeek 通过大规模专家并行（EP）与计算通信重叠（DP），大幅提升算力效率

DeepSeek-V3/R1 模型采用高度稀疏的 MoE 架构，每层专家数量众多，包含 256 个专家，但每次前向传播仅激活其中的 8 个，导致大量其他专家处于闲置状态。如果批量大小（batch size）不够大，每个专家处理的数据量会非常有限，带来计算资源利用不足、吞吐量低的问题。

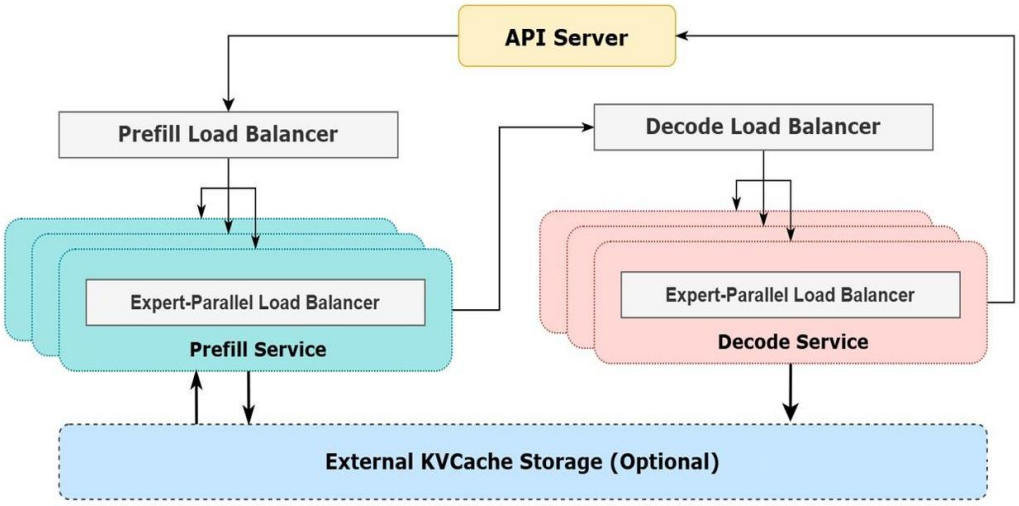
为解决以上问题，DeepSeek 采用大规模跨节点专家并行（EP）。EP 通过显存资源解耦、计算负载重构等，将专家参数分布式存储在多个 GPU 中，使得被激活的专家能够分散到不同的 GPU 进行处理，由此提升了吞吐能力、GPU 算力资源利用率也得到了提高。同时由于每个 GPU 仅处理一小部分专家，延迟也得到了降低。

图表1：DeepSeek-V3/R1 采用了预填充-解码分离架构，有效提升算力效率

阶段	部署节点数	GPU/节点	部署 GPU 数	路由专家并行度（EP）	冗余路由专家数	每 GPU 路由专家数	每 GPU 共享专家数
Prefill 预填充	4	8	32	EP32	32	9	1
Decode 解码分离	18	8	144	EP144	32	2	1

来源：DeepSeek《DeepSeek-V3/R1 推理系统概览》，国金证券研究所

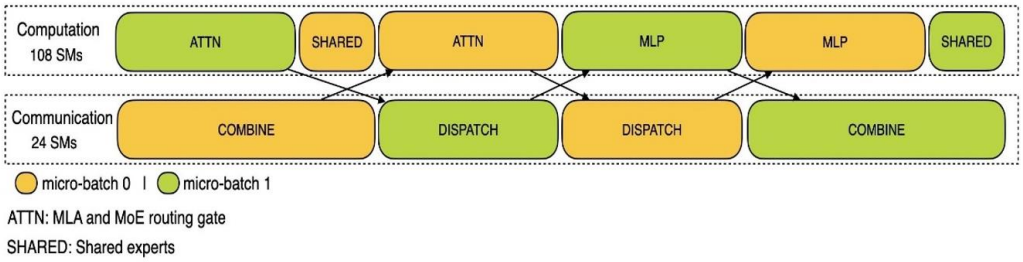
图表2：DeepSeek-V3/R1 参考架构图



来源：DeepSeek《DeepSeek-V3/R1 推理系统概览》，国金证券研究所

大规模跨节点专家并行（EP）带来较大通信开销，因此还采用计算通信重叠（DP）以缓解这一问题。在 EP 模式下，跨节点数据传输会引入额外耗时、提高通信开销。为缓解这一问题，DeepSeek 采用双批次重叠策略，通过交替处理两个批次以减少通信成本。例如在预填充阶段，DeepSeek 将一批请求分成两个微批次（micro-batch），当一个微批次正在进行计算时，另一个微批次的数据正在被传输或准备。

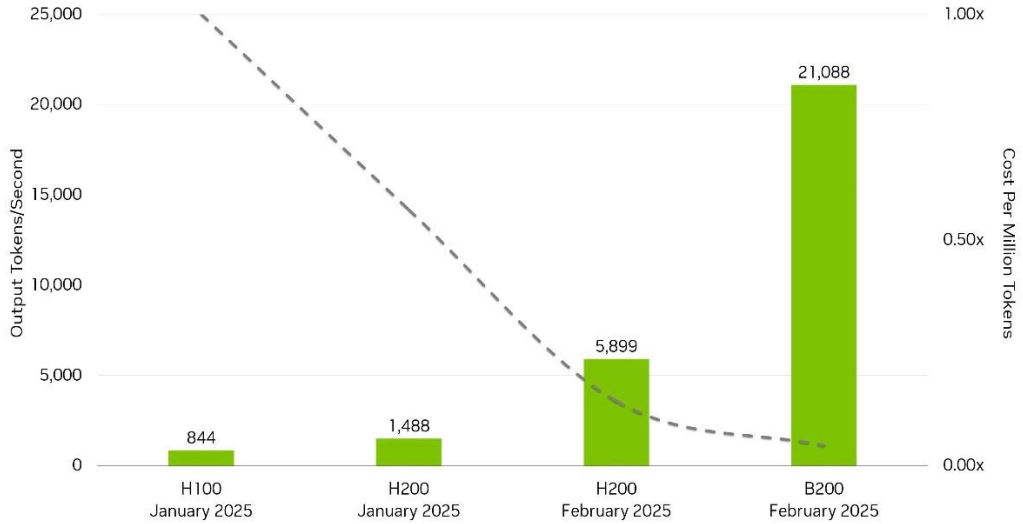
图表3：预填充阶段，将一批请求分成两个微批次



来源：DeepSeek《DeepSeek-V3/R1 推理系统概览》，国金证券研究所

通过大规模专家并行（EP）与计算通信重叠（DP），公司吞吐能力大幅提升，算力效率提高。根据 DeepSeek，对于 decode 任务，其平均每台 H800 输出吞吐约 14.8ktokens/s。作为对比，2025 年 2 月，优化后的英伟达 H200 的节点峰值输出吞吐仅 5.9ktokens/s；B200 的节点峰值输出吞吐仅 21ktokens/s。吞吐率是衡量 NLP 模型性能的核心指标，表示在单位时间内能处理的文本标记（Token）数量。DeepSeek 在使用性能低于 H200 的 H800 的前提下，吞吐能力仍然高于 H200，算力效率极高。

图表4：英伟达接入 DeepSeek-R1 后，实现 H200 和 B200 的吞吐能力提升

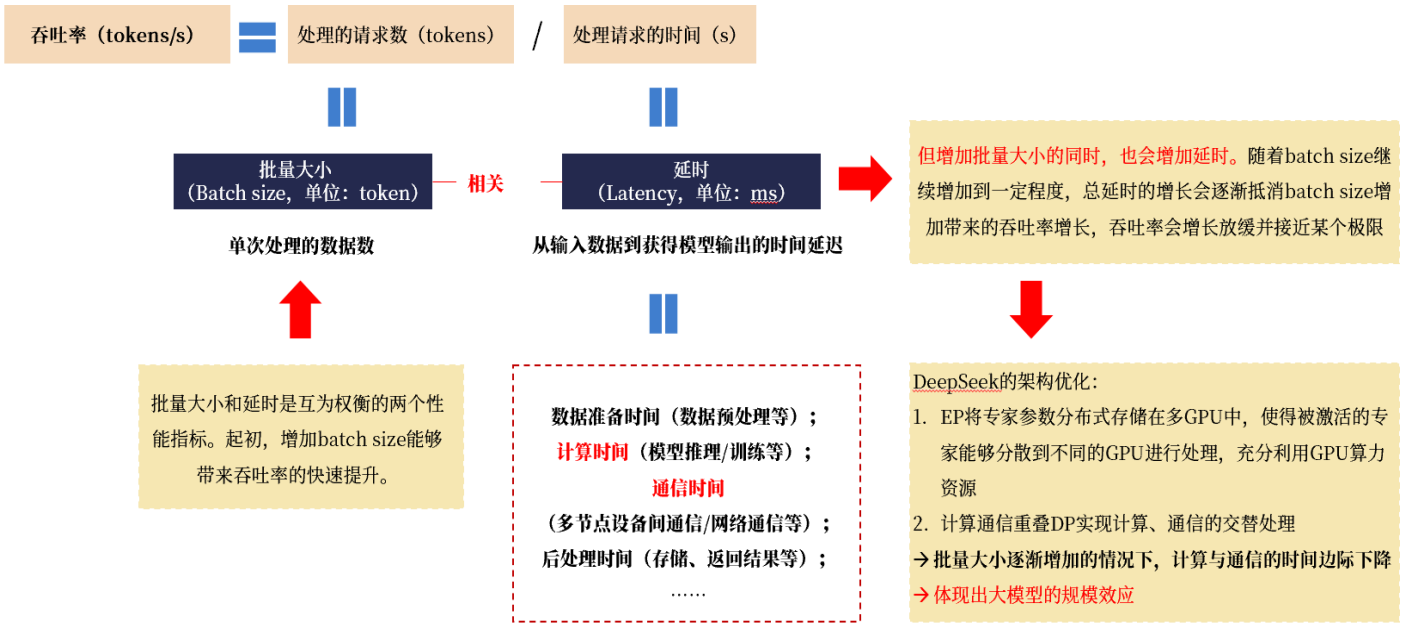


来源：Nvidia AI Developer，国金证券研究所

吞吐率取决于批量大小（batch size）与延时（latency）。批量大小和延时是互为权衡的两个性能指标。起初，增加 batch size 能够带来吞吐率的快速提升，但增加批量大小的同时，也会增加延时。随着 batch size 继续增加到一定程度，总延时的增长会逐渐抵消 batch size 增长带来的吞吐量收益，吞吐率会增长放缓并接近某个极限。

DeepSeek 通过 EP 与 DP 实现架构优化，实现了在增加批量大小的同时，延时（计算时间、通信时间）边际下降，由此吞吐率得到提升，体现出大模型的规模效应。因此大规模集群能提高算力利用率。

图表5：DeepSeek进行架构优化后，批量大小的增加将更有效地带动吞吐率的提升



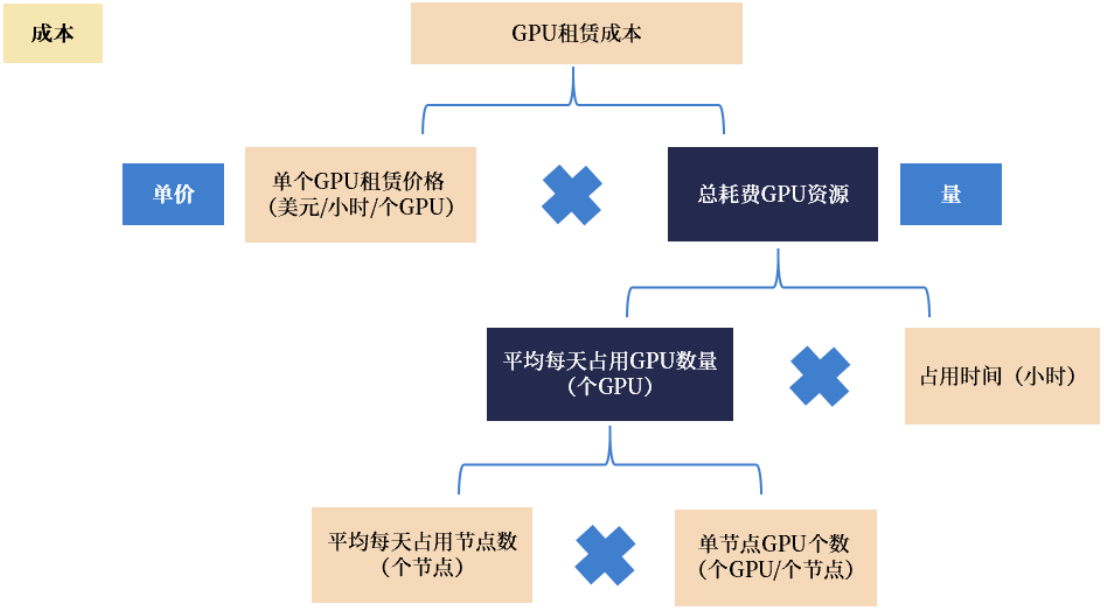
来源：DeepSeek《DeepSeek-V3/R1 推理系统概览》，壁仞科技官网，CSDN，国金证券研究所

2、参考 DeepSeek，MaaS 厂商具有盈利潜力，率先实现规模效应的云厂商将脱颖而出

2.1. DeepSeek 口径下，545%高利润率的计算详解

成本端：DeepSeek 在计算时仅考虑了 GPU 的租赁成本，约为$87,072/天。根据 DeepSeek 在知乎发布的文章《DeepSeek-V3/R1 推理系统概览》，我们可以得到以下条件：A）GPU 租赁价格:GPU 租赁成本为 2 美金/小时;B)总耗费 GPU 资源:在最近的 24 小时里,DeepSeek V3 和 R1 推理服务平均占用 226.75 个节点(每个节点为 8 个 H800),即,用到 226.75*8=1814 个 H800。GPU 运行时间即为 24 小时。由 GPU 租赁成本=单个 GPU 租赁价格*总耗费 GPU 资源，则可得到总 GPU 租赁成本。

图表6：DeepSeek 口径下，其 GPU 租赁成本测算得到$87,072/天

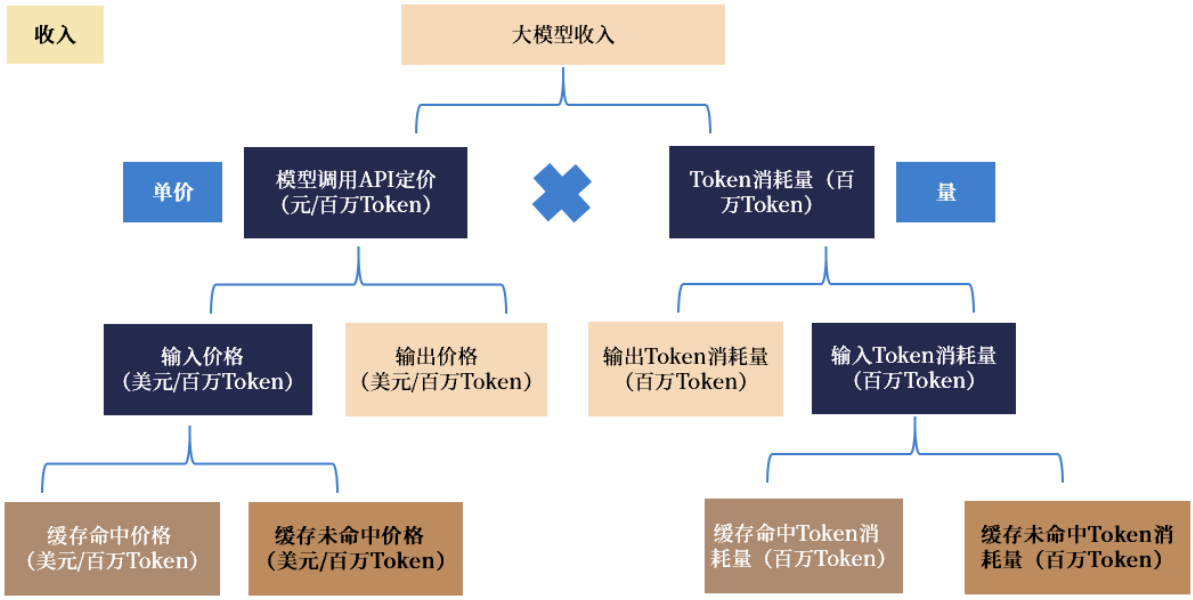


来源：DeepSeek《DeepSeek-V3/R1 推理系统概览》，国金证券研究所

收入端：所有 tokens 均根据 DeepSeek R1 的 API 定价，大模型收入约为$562,100/天。大模型带来的收入主要是指用户接入 DeepSeek 大模型的 API 而进行的费用支付，可通过模型调用 API 定价与 Tokens 消耗量相乘而得。为简化计算，DeepSeek 假设所有 tokens 全部按照 DeepSeek R1 的 API 定价计算，即，每一百万输出 tokens 定价为$2.19；每一百万输入 token（缓存命中）定价为$0.14，每一百万输入 token（缓存未命中）定价为$0.55。

在 tokens 消耗量方面，DeepSeek 披露了三个情境下的 Token 使用量：输入 token 总数为 608B，其中 342B（56.3%）为缓存命中；266B（43.7%）为缓存未命中。输出 token 总数为 168B。量价相乘，即可得到大模型收入为$562,100/天。

图表7：DeepSeek 口径下，其大模型收入测算得到约为$562,100/天



来源：DeepSeek《DeepSeek-V3/R1 推理系统概览》，国金证券研究所

利润端：公司计算得到 545%的成本利润率，对应 84.5%的收入利润率。DeepSeek 根据“AI 大模型产生的利润/GPU 租赁成本”计算得到大模型的理论成本利润率约为 545%。按照“利润率=利润/总收入”的一般财务会计口径计算，则对应 84.5%的利润率。

图表8：按利润率=利润/总收入计，DeepSeek 利润率约为 84.5%

来源：DeepSeek《DeepSeek-V3/R1 推理系统概览》，国金证券研究所

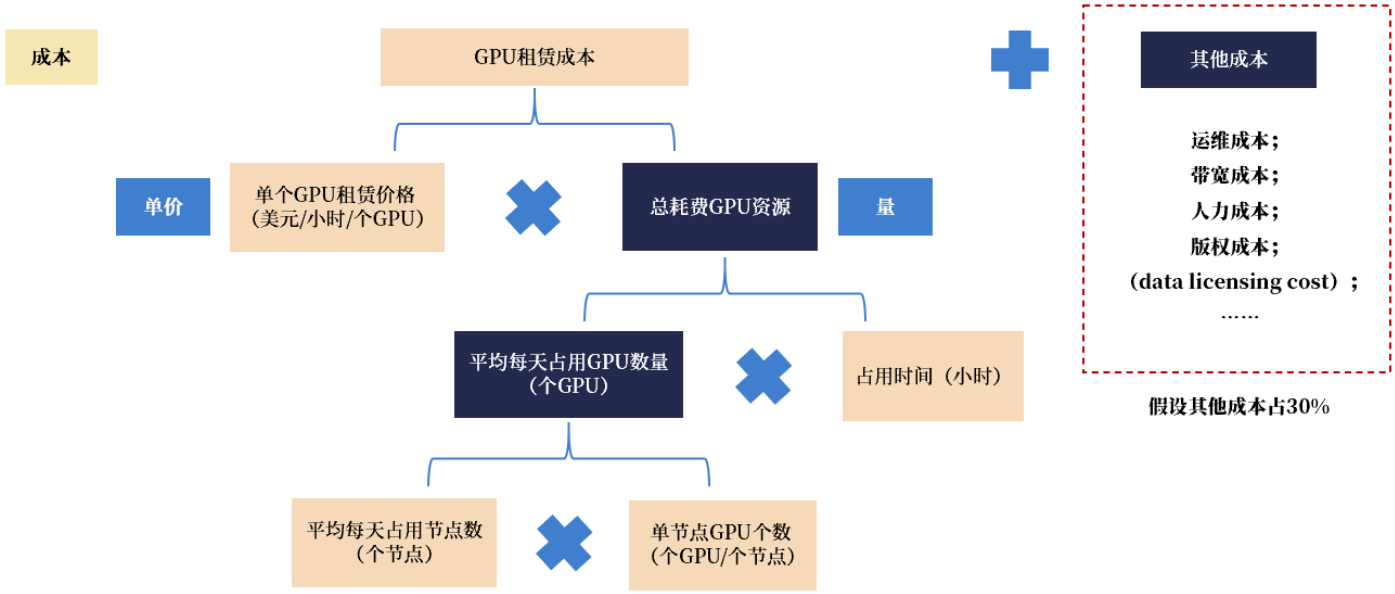
2.2. 根据实际情况调整后，DeepSeek 利润率有所降低，我们认为付费率为关键影响因素

我们根据实际情况对 DeepSeek 的收入和成本端进行了调整，以进一步分析利润率的影响因素。具体过程如下：

1）成本端调整：需考虑运维成本、带宽成本等其他经营性成本

为计算简便，DeepSeek 在成本端只计算了 GPU 的租赁价格，但未考虑到其他成本（如运维成本、带宽成本、人力成本、数据版权成本等经营性开支），由此可能会造成成本端的低估、利润率的高估。我们假设 GPU 租赁成本约占总成本的 70%，其他成本占 30%，则对应实际总成本约为$124,388.6/天。

图表9：大模型厂商的成本还包括运维、带宽、人力等其他成本



来源：DeepSeek《DeepSeek-V3/R1 推理系统概览》，国金证券研究所

2）收入端调整：需考虑 R1 定价区别、优惠定价折扣、实际付费率

为计算简便，DeepSeek 在进行利润率估算时简化了以下三个因素，因此实际大模型收入及其利润率水平或被高估。

A）R1 和 V3 的定价区别：DeepSeek 假设所有 tokens 全部按照 DeepSeek R1 的定价计算，但用于计算的 token 总数，却是 DeepSeek V3 和 R1 共同输入、输出的 token 总数。而根据 DeepSeek 官网，在标准时段内，同样情形下（缓存命中/缓存未命中/输出价格），DeepSeek V3 的价格仅为 R1 的 1/2。

B）标准时段与优惠时段的定价区别：根据 DeepSeek 官网，DeepSeek API 实行错峰优惠定价，每日优惠时段为北京时间 00:30-08:30，其余时间按照标准价格计费。而优惠时段中，DeepSeek V1 的价格仅为标准时段的 1/2，DeepSeek R1 的价格仅为标准时段的 1/4。

C）Token 调用过程中付费率：公司的统计口径包括了网页、APP 和 API 的所有负载，但 DeepSeek 的网页端和 APP 端入口均为免费，仅接入 API 的时候需要付费，因此用户付费率仅为 API 使用占比。

图表10：为计算简便，DeepSeek 在进行利润率估算时简化了三个因素

来源：DeepSeek《DeepSeek-V3/R1 推理系统概览》，国金证券研究所

我们分别对其进行假设，并纳入实际总收入的测算：

A）假设 DeepSeek V3/R1 调用需求占比分别为 35%/65%

根据 IDC 与浪潮信息发布的《2025 年中国人工智能计算力发展评估报告》，2024 年中国训练算力：推理算力约为 35%：65%。由于 DeepSeek V3 为训练大模型、R1 为推理大模型，我们假设 DeepSeek V3/R1 的需求占比与全国平均类似。由此，在原先 DeepSeek 官网对不同时段、不同模型、不同情形的定价基础上，我们根据“V3 调用单价*V3 调用需求占比+R1 调用单价*R1 调用需求占比”计算各时段、各情形的均价。

图表11：DeepSeek 官网对不同时段、不同模型、不同情形的定价

		V3	R1
标准时段价格（北京时间 08:30-00:30）	缓存命中价格	0.07	0.14
	缓存未命中价格	0.29	0.57
	输出价格	1.14	2.29
优惠时段价格（北京时间 00:30-08:30）	缓存命中价格	0.04	0.04
	缓存未命中价格	0.14	0.14
	输出价格	0.57	0.57

来源：DeepSeek 官网，国金证券研究所

图表12：假设 DeepSeek V3/R1 调用需求占比分别为 35%/65%，则得到各时段、各情形的均价

标准时段价格（北京时间 08:30-00:30）	缓存命中价格	0.12
	缓存未命中价格	0.47
	输出价格	1.89
优惠时段价格（北京时间 00:30-08:30）	缓存命中价格	0.04
	缓存未命中价格	0.14
	输出价格	0.57

来源：DeepSeek 官网，国金证券研究所

B）假设每小时输入、输出 token 数量均匀分布

根据公司官网，每日优惠时段为北京时间 00:30-08:30，即为 8 小时，剩余 16 小时为标准时段。我们假设每小时输入、输出 token 数量均匀分布，则标准时段输入 tokens 总数为单日输入 tokens 总数的 16/24，优惠时段输入 tokens 总数为单日输入 tokens 总数的 8/24。假设输入命中率为 56.3%不变，则可得到标准时段、优惠时段的输入命中、输入未命中的 tokens 总数。同理，根据公司披露，单日输出 tokens 数量为 168B，则标准时段输出 tokens 数量为 112B，优惠时段输出 token 数量为 56B。

图表13：各时段、各情形下的 tokens 调用情况

标准时段输入 token 总数（B）	**405**
-输入命中	228
-输入未命中	177
优惠时段输入 token 总数（B）	**203**
-输入命中	114
-输入未命中	89
标准时段输出 token（B）	**112**
优惠时段输出 token（B）	**56**

来源：DeepSeek《DeepSeek-V3/R1 推理系统概览》，国金证券研究所

在不考虑付费率的情况下，我们依据各时段、各情形下的 tokens 调用情况与定价相乘，得到大模型的总收入约为 37 万美金/天，对应收入利润率高达 66%。相比未调整收入、成本之前的 84.5%略有下降，但仍然体现出非常可观的利润率水平。

图表14：不考虑付费率的情况下，DeepSeek 的利润率高达 66%

大模型收入（$/天）	**370327**
标准时段收入	**321600**
-输入命中 token 数	228
-输入未命中 token 数	177
-输出 token 数	112
-输入命中单价	0.12
-输入未命中单价	0.47
-输出单价	1.89
优惠时段收入（$/天）	**48727**
-输入命中 token 数	114
-输入未命中 token 数	89
-输出 token 数	56
-输入命中单价	0.04
-输入未命中单价	0.14
-输出单价	0.57
大模型收入利润率（%）	**66%**

来源：DeepSeek《DeepSeek-V3/R1 推理系统概览》，DeepSeek 官网，国金证券研究所

C）假设付费率约为 35%

参考 ChatGPT 的收入结构，API 接口与企业端收入合计占比约 44%。根据 FutureSearch，API 收入占 OpenAI 总收入的 15%左右，针对大型企业客户的 ChatGPT Enterprise/面向中小企业的 ChatGPT Team 则分别贡献收入的 21%/8%；个人订阅用户收入占比约为 55%。因此，API 与企业端收入合计占比 44%。

由于 DeepSeek 仅 API 端口付费，网页端、APP 端均免费，而日均调用 tokens 数量是基于所有端口统计的，因此需考虑付费率对收入的影响。我们认为调用 DeepSeek API 的多为企业用户以及付费意愿较强的个人用户，参考 ChatGPT 的 API 接口与企业端口合计占比约 44%，考虑到国内付费意愿相对较低，我们假设 DeepSeek 整体付费率在 35%左右。

基于 35%的付费率，得到大模型收入约$13 万/天。

敏感性分析：我们对 GPU 租赁成本占比、API 付费率进行了敏感性分析，付费率与 GPU 租赁成本占比对公司利润率影响较大。若能将付费率提升至 40%+，则公司的利润率水平有望达 20%+。

图表15：DeepSeek 利润率敏感性分析

利润率（%）		API 付费率				
		25%	30%	35%	40%	45%
GPU 租赁成本占比	60%	-56.7%	-30.6%	-12.0%	2.0%	12.9%
	70%	-34.4%	-12.0%	4.0%	16.0%	25.4%
	80%	-17.6%	2.0%	16.0%	26.5%	34.7%

来源：DeepSeek《DeepSeek-V3/R1 推理系统概览》，DeepSeek 官网，国金证券研究所

2.3. MaaS 模式具有盈利潜力，看好能形成规模效应的公有云厂商

我们认为上述针对 DeepSeek 利润率的计算能够部分证伪市场此前认为大模型厂商商业模式无法盈利、不可持续的观点，MaaS 模式具有盈利潜力。

我们看好能实现大集群、形成高用户并发的公有云厂商，具体理由如下：

1）在第一章对 DeepSeek 底层架构优化的分析中，我们阐明了大规模集群能显著提高算力的利用率，产生规模效应。

2）规模效应能带来成本的摊薄，公有云厂商才有意愿降低定价，并吸引更多客户使用大

模型，这进而会提升客户的付费率。而伴随客户付费率的提高，公司的盈利水平也将不断提升，形成正向循环。

因此，我们看好拥有大规模集群、能形成高用户并发的公有云厂商，MaaS 盈利模式能够跑通。

3、算力需求之争：算力效率是新的 Scaling Law 方向，多模态与 AI Agent 将打开算力的成长空间

根据 DeepSeek 发布的文章，其用于单日推理服务的 H800 数量仅 1814 个，而其日活 DAU 约有 2500 万。此文引发了市场对算力通缩的担忧。

我们就 DeepSeek 的模型参数量、数据规模、峰值倍数、单卡算力、单卡利用率等关键指标进行了详细的拆解，发现 DeepSeek 低算力的原因在于：1）超高的算力效率，具体体现在单次推理激活的模型参数量、高单卡利用率；2）低峰值倍数：未设置较大的算力冗余。

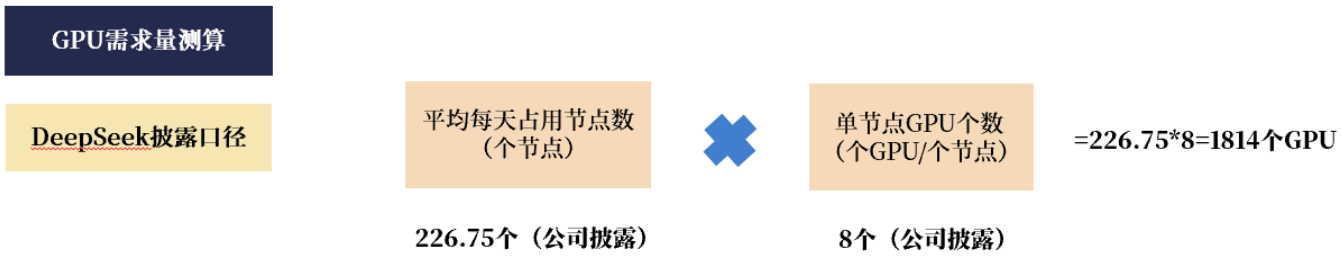
针对市场对算力通缩的担忧，我们认为：1）DeepSeek 的案例凸显了算力效率的重要性，厂商需要通过底层架构设计等提升算力效率，参数量*效率*数据规模（而非过去“参数量*数据数据规模”）才是新的 scaling law 方向。2）而从远期看，AI 落地的场景不止于聊天机器人，多品类 APP 接入 AI 大模型、多模态、AI Agent 等的出现，都将带动算力需求的提升，我们持续看好算力链。

3.1. DeepSeek 的 1814 个 GPU：算力效率的提升、算力冗余和用户使用频次的不足

DeepSeek 公布了 24 小时内用于推理服务的 H800 节点数量。根据 DeepSeek 的官方文章，在过去的 24 小时内，DeepSeek V3 和 R1 推理服务平均占用 226.75 个节点，根据单节点对应 8 个 GPU 计算，可得到总共用于推理服务的 H800 数量为 1814 个。

根据量子位披露的数据，DeepSeek 的平均日活 DAU 约为 2500 万。DeepSeek 仅用 1814 个 H800 支持了 2500 万日活用户的推理需求，引发了市场对算力需求通缩的担忧。

图表16：DeepSeek 披露其过去的 24 小时内用于推理服务的 H800 数量约为 1814 个



来源：DeepSeek《DeepSeek-V3/R1 推理系统概览》，国金证券研究所

我们基于大模型推理的算力需求、单卡算力等指标，对 DeepSeek 的算力利用情况进行了详细拆解。

1）算力需求：结合 OpenAI 的论文《Scaling Laws for Neural Language Models》，我们得到大模型推理算力需求的一般公式：推理算力需求≈2×模型参数量×数据规模×峰值倍数。

其中：A)模型参数量：参考 DeepSeek 发布的论文《DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning》，R1 每次推理仅激活 37B 参数；B）数据规模：参考 DeepSeek 披露的日均处理 token 数，换算得到 0.09 亿 tokens/秒；C）峰值倍数：我们按峰值占用节点数/平均节点数估算，得到 1.23 倍。

2）H800 单卡算力：参考 DeepSeek 开源周公布数据，H800 峰值算力达 580TFLOPS。

因此，我们可以求得 DeepSeek 对 H800 的单卡利用率高达 77%。

图表17：我们推算得到 DeepSeek 对 H800 的单卡利用率高达 77%

	数值	单位
日均 tokens 调用数量	7760.0	亿
日均总时长	86400.0	秒
数据规模	0.1	亿 tokens/秒
峰值算力倍数	1.2	倍
模型参数量	370.0	亿

	数值	单位
算力需求	0.8	EFLOPS
算力需求	**814849.1**	**TFLOPS**
H800 峰值算力	**580.0**	**TFLOPS**
H800 使用量	**1814**	**片**
H800 单卡利用率	77%	

来源：OpenAI《Scaling Laws for Neural Language Models》，DeepSeek《DeepSeek-R1：Incentivizing Reasoning Capability in LLMs via Reinforcement Learning》，DeepSeek《DeepSeek-V3/R1 推理系统概览》，DeepSeek-AI《FlashMLA》，国金证券研究所

基于上述拆解，我们提炼出影响算力需求的三个要素：

A）算力效率：具体体现在模型参数量、单卡利用率

DeepSeek 通过混合专家模型（MoE，Mixture of Experts），每次前向传播过程中，只有少数专家参与计算，而其他专家则处于闲置状态，因此，尽管 DeepSeek R1 大模型的总参数量有 671B，但单次推理激活的参数数量仅 37B，进行推理服务时仅需考虑被激活的参数数量，带动算力需求下降。

从单卡算力利用率看，DeepSeek 对 H800 的利用率高达 77%，接近峰值算力。

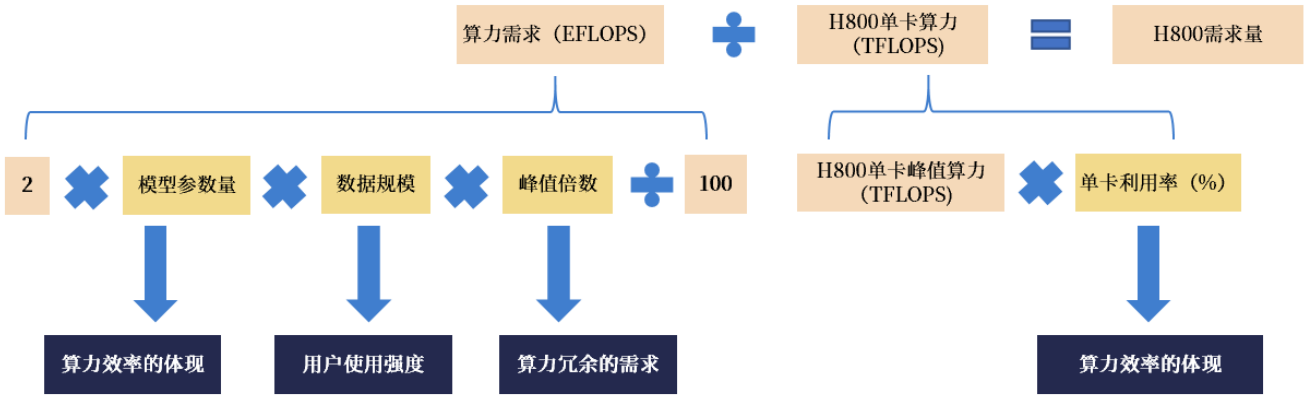
B）算力冗余需求：具体体现在峰值倍数

根据 DeepSeek 的文章进行测算，其用于推理服务的 1814 张 H800 的峰值倍数仅 1.23 倍左右。而若要应对流量突发、硬件故障或负载波动等，一般会预留更多的冗余资源比例。因此，用户在使用 DeepSeek 的过程中，常常会碰到服务器繁忙、无法使用的情况。

C）用户使用强度：具体体现在数据规模

数据规模即大模型每秒处理的 token 数目，根据 DeepSeek 披露，其日均 token 调用数约为 7760 亿，对应数据规模约为 900 万 tokens/秒。

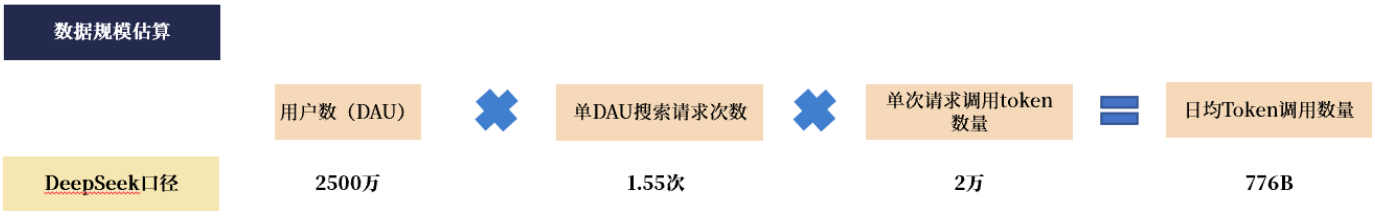
图表18：影响算力需求的因素包括算力效率、算力冗余、用户使用强度



来源：OpenAI《Scaling Laws for Neural Language Models》，国金证券研究所

我们对日均 token 调用数量进行进一步拆解，即，DAU*单 DAU 搜索请求次数*单次请求调用 tokens 数量。我们假设单次请求调用 tokens 数量约为 2 万，基于 DeepSeek2500 万的 DAU，则可以测算得到，DeepSeek 单 DAU 搜索请求次数（即，用户使用频次）仅 1.55 次。说明未来用户数、使用频次的提升均有望带来数据规模的进一步增加。

图表19：根据我们估算，当前单位 DAU 搜索请求次数仅 1.55 次



来源：量子位，硅基流动公众号，DeepSeek《DeepSeek-V3/R1 推理系统概览》，国金证券研究所

基于以上拆解，我们发现 DeepSeek 低算力的原因在于：1）超高的算力效率，具体体现在单次推理激活的模型参数量、高单卡利用率；2）低峰值倍数：未设置较大的算力冗余。

但 DeepSeek 高算力效率不等于算力通缩。我们认为算力卡的数量仍然重要，但也更考验厂商通过底层架构设计等提升算力效率的能力，参数量*效率*数据规模（而非过去“参数量*数据规模”）是新的 scaling law 方向。同时，伴随用户数、使用频次的提升，数据规模有望进一步扩大，带动算力需求提升。在下文中我们将就此展开分析。

3.2. AI Chatbot 的算力需求量估算：我们预计约 60-70 万片

DeepSeek 的大模型推理服务是 AI Chatbot 聊天机器人的典型应用，基于前文 DeepSeek 相关假设，我们估算 AI Chatbot 应用的芯片需求量约为 60-70 万片。

具体新增假设如下：

1）使用 AI 芯片假设：考虑到 H800 已经禁售，当前国内主流 AI 芯片采用的是英伟达提供的 H20，而 H20 的 FP16 峰值算力仅为 148TFLOPS。

2）单卡利用率：参考 DeepSeek 的单卡利用率，我们给予行业平均约 70%的利用率水平。

3）用户数：2025 年 1 月，ChatGPT 的日活约为 5323 万，我们假设 AI Chatbot 应用的远期用户数能达 5000 万人左右。

4）峰值倍数：为保证服务，我们设置了一定的算力冗余，假设峰值倍数为 4。

图表20：AI Chatbot 应用的 AI 芯片需求量约为 66 万片

	数值	单位
H20 峰值算力	148	TFLOPS
单卡利用率	70%	
用户数	0.5	亿人
单用户每天搜索请求次数	20.0	次
单次请求调用 tokens 数量	20000.0	个 tokens
日均 tokens 调用数量	200000.0	亿
日均总时长	86400.0	秒
数据规模	2.3	亿 tokens/秒
峰值算力倍数	4.0	倍
模型参数量	370.0	亿
算力需求	68.5	EFLOPS
算力需求	68518518.5	TFLOPS
H20 需求量	66.1	万

来源：电子发烧友网，AI 产品榜，硅基流动公众号，DeepSeek《DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning》，国金证券研究所

针对应用的用户数与单用户每天搜索请求次数，我们还进行了敏感性分析。用户数范围约在 0.3-1 亿，单用户每天搜索请求次数范围约在 10-30，对应 H20 的需求量如下图所示。

图表21：AI Chatbot 应用的算力需求量敏感性分析

H20 需求量（万）		单用户每天搜索请求次数（次）				
		10.0	15.0	20.0	25.0	30.0
用户数（亿人）	0.3	19.8	29.8	39.7	49.6	59.5
	0.4	26.5	39.7	52.9	66.1	79.4
	0.5	33.1	49.6	66.1	82.7	99.2
	0.8	52.9	79.4	105.8	132.3	158.7
	1.0	66.1	99.2	132.3	165.3	198.4

来源：国金证券研究所

3.3. 多品类 APP 接入 AI 大模型、多模态、AI Agent 原生产品带动算力需求扩容

市场担心 DeepSeek 算力效率的提高大幅降低了算力需求，当前算力芯片存量已经冗余。

我们认为基于 DeepSeek 测算出的主要是 AI Chatbot 应用的 AI 芯片需求量，多品类 APP 接入 AI 大模型、多模态、AI Agent 等 AI 原生产品将带来算力需求的进一步扩容。

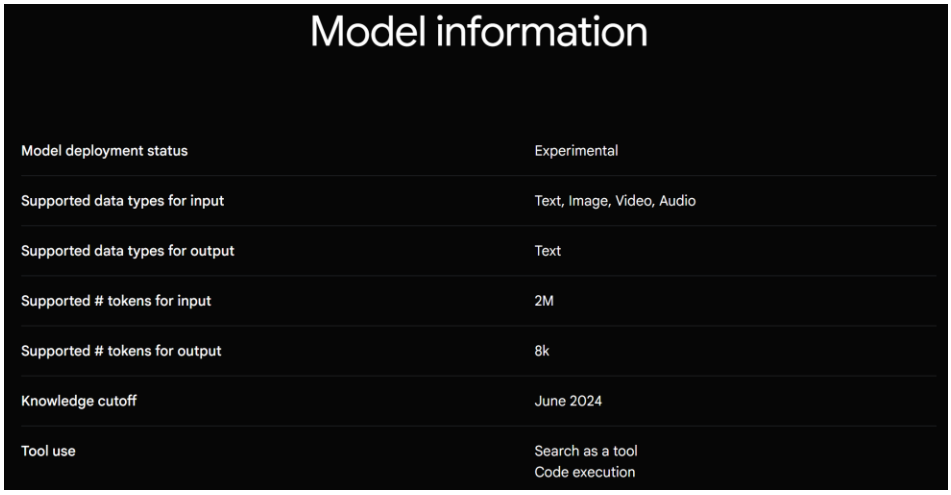
1）多品类 APP 接入 AI 大模型：现有 APP 接入 AI 大模型，带来用户数的增长

各 APP 可接入 AI 大模型（如 DeepSeek）实现应用的 AI 化；对应各 APP 的用户将在 APP 内与 AI 进行交互。因此，我们认为远期 AI 大模型的用户数也不应仅仅局限于 DeepSeek 单个 APP 的日活 DAU 数量进行估算，甚至也不应局限于单个 APP 的 DAU。从某种程度上而言，其他 APP 接入大模型本质是给大模型引流，远期用户数或可展望亿级。

2）多模态、AI Agent 带来单次请求调用 tokens 数量的大幅增加

DeepSeek 中的对话主要基于纯文字，并没有考虑到多模态大模型（除了纯文字以外，还包括图片、音频、视频等内容载体）带来的算力需求增量。而图片、音频、视频等内容将带来单次请求调用 tokens 数量的大幅增长。参考 DeepSeek 官网给出的 Token 用量计算方法，一般模型中 1 个英文字符约等于 0.3 个 token，而 1 个中文字符约等于 0.6 个 token。而根据谷歌，在 Gemini 2.0 多模态大模型中，单个图片对应 258 个 tokens，单个视频对应 263 个 tokens，单个音频则对应 32 个 tokens。最强 Gemini2.0 Pro 版本最多可支持 2M 的上下文长度，而对比 DeepSeek R1 的上下文长度最多仅支持约 64K。

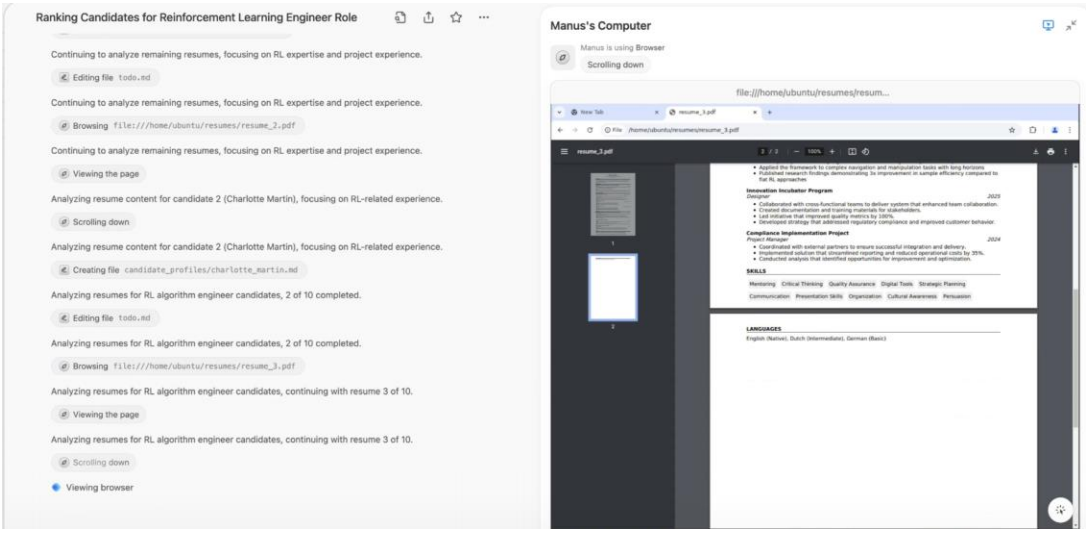
图表22：Gemini2.0 Pro 可支持 2M 的上下文长度



来源：Google Deepmind 官网，国金证券研究所

Manus 的发布引发全网关于 AI Agent 的热烈讨论，AI Agent 将带来单次请求调用 tokens 数量的大幅增加。根据智东西公众号，一个任务交给 Manus 执行后，耗费了约 24 万 tokens、3 小时，Manus 执行了 txt 文件的下载、保存、生成等；而同样的问题通过 DeepSeek 则可以秒出答案，但仅有文字回复。AI Agent 所耗费的 tokens 数量大幅增加。

图表23：Manus 能自发执行任务



来源：Manus 官网，国金证券研究所

我们估算多模态、AI Agent 的发展将带来用户单次请求调用 tokens 数量增加至 1M。如前文所述，根据 Gemini2.0 Pro，其最多可支持 2M tokens 的上下文长度；同时，AI Agent 所耗费的 tokens 数量可能是 AI Chatbot 的百倍。综合以上因素考虑，我们假设多模态、AI Agent 的发展将带来用户单次请求调用 tokens 数量增加至 1M。

基于前文的测算框架，若同样假设多模态大模型的用户数为 5000 万，单用户每天请求次数为 5 次，单次请求调用 tokens 数量提升至 1M，则对应的 H20 需求量约 800 万。

图表24：我们估算多模态、AI Agent 的发展将带动约 800 万的 H20 需求

	数值	单位
H20 峰值算力	148	TFLOPS
单卡利用率	70%	
用户数	0.5	亿人
单用户每天搜索请求次数	5.00	次
单次请求调用 tokens 数量	1000000.0	个 tokens
日均 tokens 调用数量	2500000.0	亿
日均总时长	86400.0	秒
数据规模	28.9	亿 tokens/秒
峰值算力倍数	4.0	倍
模型参数量	370.0	亿
算力需求	856.5	EFLOPS
算力需求	856481481.5	TFLOPS
H20 需求量	826.7	万

来源：电子发烧友网，AI 产品榜，Google Deepmind 官网，DeepSeek《DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning》，国金证券研究所

4、相关标的

我们认为算力需求将持续强劲，建议持续关注算力链板块。

1）我们看好能实现大集群、形成高用户并发的公有云厂商

三大运营商：云业务规模优势较为明显，具有安全可靠、IDC 资源端的优势

2024 年上半年天翼云/移动云/联通云分别实现收入 552 亿元/504 亿元/317 亿元，同比+20.4%/+19.3%/+24.3%，增长强劲。根据 IDC 数据，1H24 国内天翼云/移动云市场份额分别约为 13.2%/9.1%。结合其收入，我们估算联通云的市场份额约为 7.6%，则三大运营商合计占据国内约 30%的份额，具有较明显的规模优势。

我们认为三大运营商云的优势在于：1）提供的云服务安全可靠度高，受到政企、央国企等对安全性要求较高的客户的青睐；2）运营商的 IDC 资源辐射全国，可提供低至县域层级的本地化服务，算力服务覆盖面广，规模效应有望加强。

建议关注：中国电信、中国移动、中国联通

阿里云&腾讯云：丰富的客户资源积累、集团内其他事业部生态赋能

阿里云是国内 IaaS 龙头，根据 IDC，其 1H24 国内 IaaS 市占率约 26%，IaaS 收入 220.62 亿元。近日，阿里巴巴公告未来三年将投入 3800 亿元用于 AI 和云计算基础设施，巩固其作为全球领先的云计算供应商的地位，我们预计阿里云收入将得到大幅提振。

腾讯云位居国内 IaaS 市场的第五位，份额约 8.5%，IaaS 收入约 72.7 亿元。在阿里、字节大幅度提振资本开支的背景下，腾讯的资本开支也有望抬升，公司 AI 和云计算业务收入有望增长。

我们认为，阿里云、腾讯云等互联网厂商云业务的优势在于：1）公有云业务位居国内前列，已有丰富的客户资源积累，能形成规模效应；2）阿里、腾讯集团内其他事业部也有望接入 AI 大模型，带动云业务收入增长。

建议关注：阿里巴巴-W、腾讯控股

2）我们看好深度参与算力产业链的国产芯片、交换机厂商

寒武纪-U：AI 芯片国产替代加速，公司 AI 芯片产品性能获互联网大厂认可

AI 芯片国产替代有望加速。从需求端看，DeepSeek 带动 AI 向各行业渗透，多模态、AI Agent 打开远期算力需求空间。从供给端看，英伟达特供中国的 AI 芯片性能有所下降，同时还有禁售风险。根据彭博消息，特朗普政府正在考虑对华 H20 芯片销售实施限制。这将带来较大的缺口，有望加速 AI 芯片国产化的进程。

公司 4Q24 单季度首次实现扭亏为盈，AI 芯片产品性能得到互联网大厂认可，有望率先受益。根据公告，公司 4Q24 单季度首次实现扭亏为盈，收入/归母净利润分别约为 9.9 亿元/2.8 亿元，业绩迎来拐点。公司的智能芯片产品重点在互联网、大模型等前沿领域里，与头部客户进行了产品应用和先进技术的深度合作，当前芯片产品的实测能力、迭代预期均满足了客户的需求。我们预计伴随 AI 芯片国产替代加速，公司将率先受益。

紫光股份：以太网交换机市占率提升，高速率+CPO+全光方案是公司看点

根据 IDC、Gartner、计世资讯的相关统计数据，2020 年-2023 年，公司在中国以太网交换机市场份额分别为 35.0%、35.2%、33.8%、32.9%，持续保持市场份额第二。具体场景方面，2023 年公司在中国企业网交换机、数据中心交换机、园区交换机市场，分别以 34.2%、28.4%、36.8%的市场份额排名第二。根据 IDC 发布的最新数据，2024 年第一季度，公司在中国以太网交换机、企业网交换机、园区交换机市场，分别以 34.8%、36.5%、41.6%的市场份额排名第一，实现了市场地位的提升。

在高品质网络联接方面，为满足智算需求场景，公司推出了智算网络解决方案，全面增强网络对于多元异构算力的承载能力。同时，推出了基于 DDC 架构（分布式解耦机框）的算

力集群核心交换机 H3C S12500 AI 系列，专为 AI 算力场景设计。目前公司 800G 交换机产品也已经开始小规模发货，预计 2025 年依然有较好的上涨空间。在新技术/新方案交换机领域，公司已率先发布了 51.2T 800G CPO 硅光数据中心交换机，适用于 AIGC 集群或数据中心高性能核心交换等业务场景；新华三集团已发布“全光网络 3.0 解决方案”，高速率+CPO+全光方案是公司看点。

中兴通讯：在运营商交换机市场具备强有力的竞争力，拥有自研芯片能力

公司 51.2T 盒式交换机支持 128 个 400GE 接口，达到业界一流水平，已在互联网厂商规模商用。2024 年上半年，公司盒式交换机分别以第一名和第二名中标中国联通和中国电信集采项目；并中标中国移动 2024-2025 年数据中心交换机集采项目。根据 IDC 数据，2024 年第一季度，中兴通讯在中国以太网交换机运营商市场收入实现同比增速排名第一。同时，在数据中心交换机运营商市场领域，中兴通讯市场份额跃居第二位。

公司全资子公司中兴微电子专注于通信芯片的设计。在 5G 网络中的关键芯片及在传输承载的核心芯片，公司通过自研工艺推动产品竞争力领先。我们看好算力需求增长后，后续中兴微电子外销收入的提高。

图表25：相关标的估值（截至 2025 年 3 月 20 日收盘价）

代码	证券简称	股价（元）	EPS					PE		
			2022A	2023A	2024E	2025E	2026E	2024E	2025E	2026E
600941.SH	中国移动	105.00	5.87	6.16	6.43	6.80	7.19	18.37	15.44	14.61
601728.SH	中国电信	7.51	0.30	0.33	0.36	0.39	0.41	20.95	19.48	18.19
600050.SH	中国联通	5.71	0.23	0.26	0.28	0.34	0.38	18.70	18.13	16.39
9988.HK	阿里巴巴-W	135.50	2.84	3.43	3.91	6.88	7.54	16.27	18.19	16.60
0700.HK	腾讯控股	519.50	19.34	11.89	20.49	22.85	25.55	18.36	21.00	18.78
688256.SH	寒武纪-U	703.63	-3.13	-2.04	-1.06	0.11	1.17	-619.79	6324.66	599.95
000938.SZ	紫光股份	28.69	0.75	0.74	0.81	1.03	1.23	35.31	27.83	23.34
000063.SZ	中兴通讯	36.20	1.71	1.95	1.76	1.99	2.32	22.94	19.4	16.69

来源：wind，国金证券研究所

注：盈利预测中中国移动、中国联通、中兴通讯采用国金预测数据，其余采用 wind 一致预期，中国移动、中国联通、阿里巴巴-W、腾讯控股、中兴通讯 2024 年已出业绩，采用实际数据

5、风险提示

多模态大模型、AI Agent 落地不及预期的风险。我们认为，原有 APP 接入 AI 大模型、多模态大模型、AI Agent 能大幅提振算力需求。但如果 AI 在上述应用中渗透不及预期，算力需求提振幅度或有所放缓。

芯片供应不足的风险。英伟达芯片存在禁售的风险，国内芯片行业正在快速发展，但在短时间内将现有 AI 芯片全部切换为国产芯片作为替代方案的可行性较低。若未来国际经济贸易形势出现重大不利变化，行业可能面临芯片供应不足的风险。

客户对公有云接受度不及预期的风险。我们认为大规模集群能带来规模效应，提升云厂商的盈利能力，因此看好公有云厂商。但部分客户可能处于数据安全等因素的考量，不愿意采用公有云，影响公有云厂商的业绩。

行业竞争加剧的风险。随着对云服务需求不断释放，行业内参与者可能会大幅增加。业内各公司可能面临价格竞争、客户资源竞争等压力，可能存在竞争加剧导致各公司盈利能力下降的风险。

行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。

特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告版权归“国金证券股份有限公司”（以下简称“国金证券”）所有，未经事先书面授权，任何机构和个人均不得以任何方式对本报告的任何部分制作任何形式的复制、转发、转载、引用、修改、仿制、刊发，或以任何侵犯本公司版权的其他方式使用。经过书面授权的引用、刊发，需注明出处为“国金证券股份有限公司”，且不得对本报告进行任何有悖原意的删节和修改。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

此报告仅限于中国境内使用。国金证券版权所有，保留一切权利。

上海
电话：021-80234211
邮箱：researchsh@gjzq.com.cn
邮编：201204
地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 5 楼

北京
电话：010-85950438
邮箱：researchbj@gjzq.com.cn
邮编：100005
地址：北京市东城区建内大街 26 号新闻大厦 8 层南侧

深圳
电话：0755-86695353
邮箱：researchsz@gjzq.com.cn
邮编：518000
地址：深圳市福田区金田路 2028 号皇岗商务中心 18 楼 1806

【小程序】
国金证券研究服务

【公众号】
国金证券研究