通信

DeepSeek：模型效率的革命，算力基建的新起点

【我们团队对于 DeepSeek 理解，可以用一个非常巧妙的比喻来概括：内燃机效率的提高，一定会带来更大的石油消耗】

DeepSeek 做了什么：

【1】验证了新的 Scaling Law：在过去的几年间，AI 模型的增长主要依靠于预训练阶段的规模堆积，通过对人类已有数据的不断积累，从而增加模型的规模和能力。但随着训练耗尽了人类已有数据，叠加合成数据之路所产生的幻觉和边际收益递减问题，预训练阶段的 Scaling Law 逐渐放缓。2024 年以来，基于强化学习的新 Scaling Law 成为了行业重点方向，先后涌现出了 GPT-o1，Deepseek R1 等优秀模型。RL 这种新的 AI 增长曲线，在初期展现出了非常高的投入性价比，这也是 Deepseek V3 快速进化到 R1，并展现出强大能力的根本原因。当下 RL 依然是基于传统的预训练模型，在推理阶段加入 RL 使得模型面对理科问题时的推理技能更加强大。展望未来，随着 RL 算法在预训练阶段逐渐取代自回归算法，同时使用更强大的算力和更多的数据，基于 RL 算法训练模型生成思维链，将共同构成 AI 新的 Scaling Law，**算法创新与算力堆叠在这条曲线上将一起狂奔，AI 的能力边界将迎来新一轮的扩张。**

【2】极致的工程优化：DeepSeek 的真正创新之处在于极致的工程优化，主要依靠了如键值缓存，创新型的 MoE 架构，以及基于 PTX 层汇编语言实现对于英伟达 GPU 效率的极致压榨等等，通过这些创新性的工程优化，使得 DeepSeek 打破了西方模型公司固有的降本周期曲线，能够用更低的价格，来提供接近头部模型的使用体验，提高了全球算力的使用效率。

【3】慷慨的开源：与西方以 OAI 和 Anthropic 为代表模型巨头，逐渐走向闭源的模型商业理念不同，DeepSeek 将自身创新性原理和模型进行了开源，一方面使得全球模型厂商能够共享新的工程方法带来的性价比提升，另一方面使得全球用户能够本地或者通过公有云部署，从而避免高溢价，这也是 Deepseek 在全球开发者中获得如此高殊荣的核心原因。

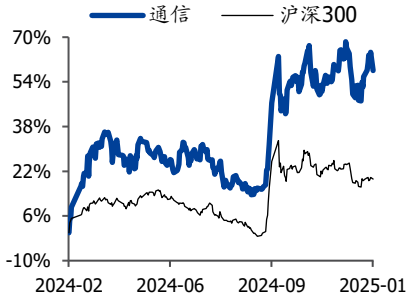
DeepSeek 对于英伟达的影响：

Deepseek 的出现，让全世界看到 AGI 的实现又更近了一步。我们认为，海外在算力领域的布局不会因为 deepseek 有所放缓，相反，因为 deepseek 的出现，将会给全球科技巨头进一步上紧发条，海外科技巨头们将进一步加码在算力领域的布局。具体而言，一方面可能会进一步加大英伟达 GPU 采购力度，另一方面也会加紧推进自研 ASIC 方案的进度。此外，美国政府可能会进一步加紧芯片出口限制，试图在算力层面上进行最后的封锁，以限制其他国家地区的 AI 发展，维护其所谓的 AI 领先地位。

对于英伟达而言，我们认为，deepseek 的阶段性胜利，将继续推动算力市场的整体需求，长期的天花板进一步被打开，我们不认为英伟达

增持（维持）

行业走势



作者

分析师 宋嘉吉
执业证书编号：S0680519010002
邮箱：songjiaji@gszq.com
分析师 黄瀚
执业证书编号：S0680519050002
邮箱：huanghan@gszq.com
分析师 邵帅
执业证书编号：S0680522120003
邮箱：shaoshuai@gszq.com

相关研究

1、《通信：如何丈量这一轮 CPO 的高度——OIO》2025-01-27
2、《通信：母线——AI 大功率集群需求的新解》2025-01-25
3、《通信：CPO：开启光学的新画卷》2025-01-19

的需求和订单会出现任何大幅度的影响，相反，一方面可能会进一步超预期，另一方面英伟达可能也会将其下一代产品的迭代速度加速，性能有望实现重大突破的极具创新型技术路线（如 CPO、OIO 等），我们认为有望继续提速。我们认为英伟达的股价下跌并非反应的是需求的变化，而更多应是担心可能发生的进一步加紧的芯片出口限制而带来的不确定性。

DeepSeek 对于中国算力投资的影响

【对于国产算力】：DeepSeek 推出后，美国迫于竞争压力，对于中国进一步加强算力制裁的呼声愈发强大。同时由于 DeepSeek 开源和低成本特性，国产显卡用于推理的性价比和 ROI 急剧上升，至此，基于中芯国际的制造能力，芯片厂商的设计能力，Deepseek 模型的应用能力三者合一，中国算力自主可控的飞轮开始旋转。

【对于边缘侧】：DeepSeek 对于边缘侧的拉动主要有两点，第一是云端先进模型的价格降低，使得 AI 应用，AI 硬件的使用和开发运营成本降低，这将促进 AI 硬件的放量。第二是 DeepSeek 对于小模型优化的能力，此次 Deepseek 也基于 Qwen 与 Llama 发布了部分优化后的小模型，能力相较于此前小模型有一定提升。我们认为，随着 RL 这一新增长曲线的不断放量，叠加边缘侧算力的提升，边缘推理将加速变为现实，边缘算力有望加速放量。

Deepseek 对于全球算力投资的影响：

1、GPU：我们认为 deepseek 对 GPU 的影响是积极的，因为目前无论从训练还是推理的角度，GPU 的优势的非常明显，算力长期来看依旧将继续指数级增长，我们依旧看好英伟达在 AI 基础设施领域的确定性优势，同时看好其在能够大幅度提升性能的新技术领域的加速，看好 CPO、OIO 的加速落地，建议关注：天孚通信、康宁、太辰光等。

2、ASIC：强化学习对推理的积极影响，可能会加速大模型团队及云厂商对于推理以及自研算力的需求和进展，最终市场取决于单位算力成本的比较，目前来看英伟达在推理领域依旧有着明显的优势，但云厂商可能会继续冒着支出更多的代价，坚定推进自研算力。利好云厂商的 ASIC 合作伙伴，以及其配套产业链：博通、amd、mrvl、arista、cls、中际旭创、新易盛、cohr、lite 等。

3、通信：deepseek 对通信的影响比较综合，通常训练集群通信要求相较于同等规模的推理集群更高，但考虑到性价比问题，目前众多训练集群也兼顾了推荐算法、推理、科研、金融交易等其他功能，我们认为训练集群依然将坚定的向着规模更大的方向发展。同时，推理需求需要考虑规模效应，预计云厂商主导的推理市场依旧将占主导地位，端侧目前算力有限，不构成太大影响，预计推理集群也将延续过去发展趋势。综合来看，无论是训练还是推理，依旧将使用到更多的高端交换机和光模块，我们倾向于推荐产品有优势、在手订单充足、产能扩展顺利的标的，建议关注：博通、arista、cls、中际旭创、新易盛、cohr、lite、ciena、德科立。

4、能源：海外能源相关配套差，我们预计 deepseek 的出现，将坚定海外科技巨头加速加码布局基础设施的决心，能源领域继续看好整个方向，建议关注：SMR、OKLO、威腾电气，电工合金。

总结来看，对于节后 A 股的思考：

春节假期中，除了 DeepSeek 之外，还发生了北美关税落地，芯片制裁进一步加严等事件。结合市场对于 DeepSeek 的边际变化的认知更多反应在应用端的快速放量上，因此节后主要的进攻方向主要聚焦在端侧以及国产替代链上，而从中长期来看，通信、能源等全球算力配套产业链也将迎来全新的发展期。

节后方向：
端侧：物联网模组龙头，美格智能，移远通信，广和通。
国产替代：寒武纪-U，中芯国际，中兴通讯。
建议关注：
算力——
光通信：中际旭创、新易盛、天孚通信、太辰光、腾景科技、光库科技、光迅科技、德科立、联特科技、华工科技、源杰科技、剑桥科技、铭普光磁。**铜链接：**沃尔核材、精达股份。**算力设备：**中兴通讯、紫光股份、锐捷网络、盛科通信、菲菱科思、工业富联、沪电股份、寒武纪、海光信息。**液冷：**英维克、申菱环境、高澜股份。**边缘算力承载平台：**美格智能、广和通、移远通信。**卫星通信：**中国卫通、中国卫星、震有科技、海格通信。**IDC：**润泽科技、光环新网、奥飞数据、科华数据、润建股份。
数据要素——

运营商：中国电信、中国移动、中国联通。**数据可视化：**浩瀚深度、恒为科技、中新赛克。

风险提示：AI 发展不及预期，算力需求不及预期，市场竞争风险。

重点标的

股票代码	股票名称	投资评级	EPS（元）				PE			
			2023A	2024E	2025E	2026E	2023A	2024E	2025E	2026E
300308.SZ	中际旭创	买入	1.94	4.99	8.73	11.58	87.70	23.03	13.16	9.91
300502.SZ	新易盛	买入	0.97	4.13	8.80	11.98	131.40	30.44	14.29	10.49
300394.SZ	天孚通信	买入	1.32	2.42	3.78	4.82	76.50	41.61	26.68	20.92
002463.SZ	沪电股份	买入	0.79	1.25	1.63	1.96	41.30	33.16	25.46	21.24
002837.SZ	英维克	买入	0.47	0.72	0.94	1.21	67.90	57.25	43.51	33.77

资料来源：Wind，国盛证券研究所

内容目录

1. 投资策略：DeepSeek—模型效率的革命，算力基建的新起点 5
2. 行情回顾：通信板块下跌，光通信表现最优 6
3. 春节专题 1：DeepSeek—模型效率的革命，算力基建的新起点 7
4. 春节专题 2：如何看待 DeepSeek 对英伟达以及算力产业链的影响 11
5. 英伟达、亚马逊、微软同日接入 DeepSeek 12
6. OpenAI 上线 o3-mini，首次向 ChatGPT 免费用户开放推理模型 13
7. 风险提示 13

图表目录

图表 1：通信板块上涨，细分板块中光通信表现相对最优 6
图表 2：本周拓尔思领涨通信行业 6
图表 3：GRPO 算法 7
图表 4：Deepseek 使用 PTX 语言 8
图表 5：Deepseek V3 采用 MoE 架构 8
图表 6：Groq 当下仅支持基于 Llama 优化的 70B 模型 9
图表 7：Deepseek 对于部分开源小模型进行了优化 10

1. 投资策略：DeepSeek—模型效率的革命，算力基建的新起点

本周建议关注：

算力——

光通信：中际旭创、新易盛、天孚通信、太辰光、腾景科技、光库科技、光迅科技、德科立、联特科技、华工科技、源杰科技、剑桥科技、铭普光磁。

铜链接：沃尔核材、精达股份。

算力设备：中兴通讯、紫光股份、锐捷网络、盛科通信、菲菱科思、工业富联、沪电股份、寒武纪、海光信息。

液冷：英维克、申菱环境、高澜股份。

边缘算力承载平台：美格智能、广和通、移远通信。

卫星通信：中国卫通、中国卫星、震有科技、海格通信。

IDC：润泽科技、光环新网、奥飞数据、科华数据、润建股份。

数据要素——

运营商：中国电信、中国移动、中国联通。

数据可视化：浩瀚深度、恒为科技、中新赛克。

本周观点变化：

本周受 Deepseek 出圈影响，国内外算力市场均出现大幅波动。国内多支相关个股跌停，海外 AI 基础设施板块巨震后，英伟达走势最弱，而博通、arista、mrvl、cls、lite、cohr、credo 等代表的非 nv 算力产业链均出现不同程度的回暖。我们认为，节后主要的进攻方向主要聚焦在端侧以及国产替代链上，而从中长期来看，通信、能源等全球算力配套产业链也将迎来全新的发展期。

2. 行情回顾：通信板块下跌，光通信表现最优

本周（2025 年 1 月 27 日）上证综指收于 3250.60 点。各行情指标从好到坏依次为：上证综指>沪深 300>万得全 A>万得全 A(除金融，石油石化)>中小板综>创业板综。通信板块下跌，表现劣于上证综指。

图表1：通信板块上涨，细分板块中光通信表现相对最优

指数	涨跌幅度
上证综指	-0.1%
沪深 300	-0.4%
万得全 A	-0.88%
万得全 A(除金融，石油石化)	-0.99%
中小板综	-1.00%
创业板综	-2.3%
国盛通信行业指数	-3.8%
国盛运营商指数	1.8%
国盛卫星通信导航指数	-2.1%
国盛物联网指数	-2.7%
国盛移动互联指数	-3.0%
国盛区块链指数	-3.23%
国盛通信设备指数	-3.25%
国盛云计算指数	-3.30%
国盛量子通信指数	-4.5%
国盛光通信指数	-7.9%

资料来源：Wind，国盛证券研究所

从细分行业指数看；运营商上涨 1.8%；卫星通信导航、物联网、移动互联、区块链、通信设备、云计算分别下跌 2.1%、2.7%、3.0%、3.2%、3.25%、3.3%，表现优于通信行业平均水平；量子通信、光通信分别下跌 4.5%、7.9%。

本周，受益于 AI 智能体概念，拓尔思上涨 17.794%，领涨版块。三变科技上涨 9.984%；浙江东方上涨 9.935%；受益于家庭医生概念，ST 易联众上涨 5.502%；受益于 AI 智能体概念；*ST 信通上涨 5.054%。

图表2：本周拓尔思领涨通信行业

涨跌幅前五名				涨跌幅后五名			
证券代码	证券名称	涨跌幅（%）	成交量（万手）	证券代码	证券名称	涨跌幅（%）	成交量（万手）
300229.SZ	拓尔思	17.794	185.56	300068.SZ	南都电源	-14.13	91.82
002112.SZ	三变科技	9.984	54.66	300394.SZ	天孚通信	-11.59	32.46
600120.SH	浙江东方	9.935	25.08	300399.SZ	天利科技	-10.56	21.17
300096.SZ	ST 易联众	5.502	10.12	300383.SZ	光环新网	-10.35	129.05
600289.SH	*ST 信通	5.054	1.38	603118.SH	共进股份	-9.97	65.00

资料来源：Wind，国盛证券研究所

3. 春节专题 1：DeepSeek—模型效率的革命，算力基建的新起点

【核心的一句话：内燃机效率的提升，将带来更多的石油消耗】

我们在首页中已经较为详细的阐述了 DeepSeek 给全球模型界带来的趋势性变化，在基于 RL 的新增长曲线，开源带来的全球使用部署热潮下，我们认为算力将在训练与推理的新一轮增长中迎来新的繁荣。

首页中提到，DeepSeek 主要的创新来自于工程实现优化，其中包括了创新性 MOE，PTX 语言等等等等。我们在正文中就一些方面进行简单展开。

在训练原理层面，DeepSeek-R1-Zero 模型创新性的跳过了 SFT 阶段，仅仅使用 RL，使得训练过程进一步简化。DeepSeek-R1-Zero 创新性的引入了 GRPO 算法，使得模型冷启动时能够跳过 supervised data，从而能够节省训练成本。

图表3：GRPO 算法

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right),$$

$$\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1,$$

资料来源：Deepseek，国盛证券研究所

第二，通过使用 PTX 语言，使得在训练与推理过程中，对于英伟达 GPU 的性能利用更为高效。英伟达的 Cuda 体系大体上可以分为三层，第一层为 Cuda 层，第二层称为 PTX 层，第三层则为机器码，对应到我们传统编程语言中的“高级语言、汇编语言、机器码”在开发时，Cuda 层可以调用函数，利用已有算力，简化流程，因此更加简单，但是无法实现对于 GPU 的精确控制，PTX 则可以深入源头，对于 GPU 实现更加精确的控制，从而有效提高算力的利用效率。在训练 V3 模型时，DeepSeek 对英伟达 H800 GPU 进行了重新配置：在 132 个流处理器多核中，划分出 20 个用于服务器间通信，主要用于数据压缩和解压缩，以突破处理器的连接限制、提升事务处理速度。

图表4：Deepseek 使用 PTX 语言

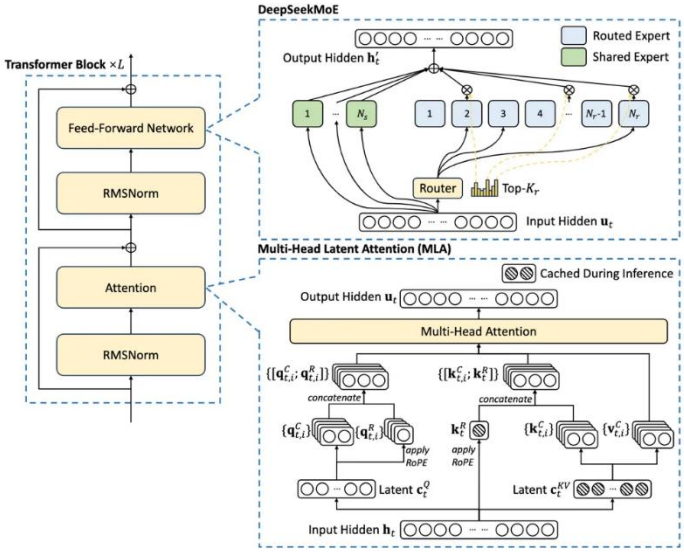
In detail, we employ the warp specialization technique (Bauer et al., 2014) and partition 20 SMs into 10 communication channels. During the dispatching process, (1) IB sending, (2) IB-to-NVLink forwarding, and (3) NVLink receiving are handled by respective warps. The number of warps allocated to each communication task is dynamically adjusted according to the actual workload across all SMs. Similarly, during the combining process, (1) NVLink sending, (2) NVLink-to-IB forwarding and accumulation, and (3) IB receiving and accumulation are also handled by dynamically adjusted warps. In addition, both dispatching and combining kernels overlap with the computation stream, so we also consider their impact on other SM computation kernels. Specifically, we employ customized PTX (Parallel Thread Execution) instructions and auto-tune the communication chunk size, which significantly reduces the use of the L2 cache and the interference to other SMs.

资料来源：Deepseek，国盛证券研究所

第三，继续保留了 DeepSeek V3 优秀的 MOE 架构，使得推理时能够更加精准的调用对应的专家参数，进一步提高了模型的性价比。DeepSeek-V3 的设计理念是，通过智能地选择激活哪些参数，而不是盲目地激活所有参数，从而实现了在有限的计算资源下，实现最优的表现。这种方法不仅提高了计算效率，也使得 DeepSeek-V3 能够在处理复杂任务时，表现出更高的性能。

在此基础上，DeepSeek-V3 的 MoE 设计还具有出色的可扩展性。这一设计通过整合不同领域的“专家”，实现了模型的灵活扩展，无需将所有模型紧密集成在一起。通过这种模块化的设计，DeepSeek-V3 能轻松地进行规模扩展，同时还能灵活地适应新的需求和挑战。这种模块化的设计方式，使得 DeepSeek-V3 能够根据需要，快速增加或减少“专家”，以适应不同的任务和环境。

图表5：Deepseek V3 采用 MoE 架构



资料来源：Deepseek，国盛证券研究所

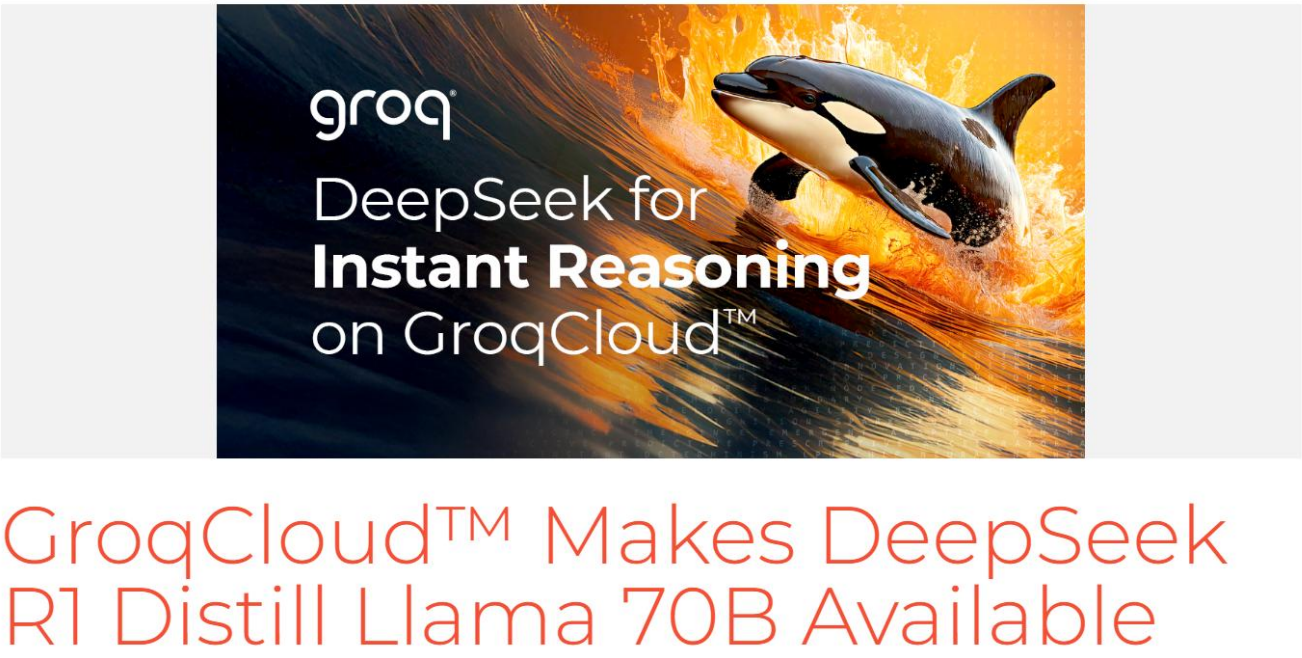
第二部分，我们对于摘要中关对于 A 股投资部分进行展开。我们认为，在 DeepSeek 带来的全球算力新周期下，A 股中的算力板块将会迎来新一轮的上行大 Beta。其中国产算力，边缘侧算力，算力基建板块是我们建议优先关注的方向。

从国产算力板块来说，在 DeepSeek 推出后，我们认为，由中芯国际的制造能力，设计公司的计算能力，以及 Deepseek 为代表的新一代开源模型带来的商用能力，三者共同构成国产算力的飞轮已经完全建立。

同时，在 DeepSeek 发布后，北美对于加码芯片封锁的意愿愈发强烈。Anthropic CEO 发文强调必须加强封锁。长期趋势下，算力国产化比例提升是中美 AI 军备下的必然趋势，而 DeepSeek 则加速了这一进程。

我们认为，从国内来看，GPGPU 依然是最适合国内需求的芯片构型。一方面 DeepSeek 采用 PTX 汇编语言，先前 ASIC 多只对英伟达的 cuda 层做了一站式的 complier 与 assmbler 软件，现在兼容 PTX 语言需要更长的适配时间，从北美当下情况来看，类似 Groq 等主流 ASIC 云都只上线了 70B 版本的蒸馏模型，迟迟未能部署 671B 的完整版 R1，这也一定程度上体现了 ASIC 对于新范式下模型的适配难度。而 GPGPU 由于天生构型与英伟达类似，对于相对应的语言转变有更加灵活的适应能力。另一方面，DeepSeek 的出现，也代表了算法变化成为了大模型进化路上的关键一环，而对于不同算法的灵活适应，也是 GPGPU 的核心优势。

图表6：Groq 当下仅支持基于 Llama 优化的 70B 模型



资料来源：Groq，国盛证券研究所

对于边缘侧来说，Deepseek 带来的变化主要有两点，第一是云端先进模型的价格降低，使得 AI 应用，AI 硬件的使用和开发运营成本降低，这将促进 AI 硬件的放量。第二是 DeepSeek 对于小模型优化的优化能力。此次 Deepseek 也基于 Qwen 与 Llama 发布了部分优化后的小模型，能力相较于此前小模型有一定提升。我们认为，随着 RL 这一新增长曲线的不断放量，叠加边缘侧算力的提升，边缘推理将加速变为现实，边缘算力有望加速放量。

图表7：Deepseek 对于部分开源小模型进行了优化

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

Table 5 | Comparison of DeepSeek-R1 distilled models and other comparable models on reasoning-related benchmarks.

资料来源：Deepseek，国盛证券研究所

4. 春节专题 2: 如何看待 DeepSeek 对英伟达以及算力产业链的影响

Deepseek 自全球出圈后，美股算力板块迎来巨震，AI 基础设施板块中，英伟达走势最弱，而博通、arista、mrvl、cls、lite、cohr、credo 等代表的非 nv 算力产业链均出现不同程度的回暖。我们发现市场普遍担心由于训练成本的大幅度下降，导致训练相关算力未来需求可能会出现疲软，而英伟达在训练领域优势最为明显，这无疑削弱了英伟达的竞争力。另一方面，随着强化学习在推理领域的 scaling law 效应的进一步体现，市场可能更加关注推理算力未来的增量，而推理相较于训练要求会有所降低，市场的选择可能会更多，这进一步加剧了市场对英伟达的担心。

但更深层次，我们应该注意到，随着 deepseek 的开源，众多团队迅速根据其论文复现成功，对全行业水平的提升贡献巨大，以 openai 为首的闭源模型面临了开源模型最直接的冲击：用户大量涌入开源模型 app，开发者争先尝试开源模型，云厂商迅速在公有云平台部署 deepseek 的开源模型，以此来继续抢占用户。Deepseek 的重磅开源，某种意义上实现了技术平权，而竞争对手们也会迅速参考其优秀的方法思路去优化自身模型，驱动了全球大模型的进步，对社会的影响巨大深远。

我们应该意识到，大模型的差距正在缩小，公开论文、开源模型以及人才的流动，使得好的工程方法和创意思考在不同的组织内流动并实施落地，未来开源和闭源之间的竞争会更加激烈，token 的价格会进一步下降，用户则充分受益于竞争，能够以更低的成本甚至是 0 成本获取到高质量模型，进而继续拉动 token 的需求量。Deepseek 自出圈后用户量暴增导致服务经常性崩溃，云厂商基于其提前部署的算力，迅速的抢占流量和用户，应用的持续升级，将进一步拉动需求的指数级放量，可以预见算力依然将持续成为最核心的竞争和门槛。

我们从本周微软和 meta 的财报可以看到，云厂商在 AI 基础设施的投入上没有任何放缓的迹象，相反，基于算力建立的壁垒甚至正在变得更高:

1、Deepseek 的算力集群，放在世界范围内也算比较领先的，如果没有超前的眼光部署算力基础设施，就不会有 deepseek。

2、模型层面的差距变小，将进一步推动巨头在算力领域的竞争，谁先抢占到最优质的算力卡，谁先搭建更大效率更高的超算集群，谁就更有可能领先竞争对手。

3、推理需求正在变得更大，拥有更高效率的推理算力的厂商，能更好的满足用户指数级爆发的需求，就会更有先机赢在未来。

deepseek 的出现，让全世界看到 AGI 的实现又更近了一步。我们认为，海外在算力领域的布局不会因为 deepseek 有所放缓，相反，因为 deepseek 的出现，将会给全球科技巨头进一步上紧发条，海外科技巨头们将进一步加码在算力领域的布局。具体而言，一方面可能会进一步加大英伟达 GPU 采购力度，另一方面也会加紧推进自研 ASIC 方案的进度。此外，美国政府可能会进一步加紧芯片出口限制，试图在算力层面上进行最后的封锁，以限制其他国家地区的 AI 发展，维护其所谓的 AI 领先地位。

对于英伟达而言，我们认为，deepseek 的阶段性胜利，将继续推动算力市场的整体需求，长期的天花板进一步被打开，我们不认为英伟达的需求和订单会出现任何大幅度的影响，相反，一方面可能会进一步超预期，另一方面英伟达可能也会将其下一代产品的迭代速度加速，性能有望实现重大突破的极具创新型技术路线（如 CPO、OIO 等），我们认为有望继续提速。我们认为英伟达的股价下跌并非反应的是需求的变化，而更多应是担心可能发生的进一步加紧的芯片出口限制而带来的不确定性。

具体到整个算力产业链，我们认为 deepseek 带来的影响整体偏积极，其中

1、GPU：我们认为 deepseek 对 GPU 的影响是积极的，因为目前无论从训练还是推理的角度，GPU 的优势的非常明显，算力长期来看依旧将继续指数级增长，我们依旧看好英伟达在 AI 基础设施领域的确定性优势，同时看好其在能够大幅度提升性能的新技术领域的加速，看好 CPO、OIO 的加速落地，建议关注：天孚通信、康宁、太辰光等。

2、ASIC：强化学习对推理的积极影响，可能会加速大模型团队及云厂商对于推理以及自研算力的需求和进展，最终市场取决于单位算力成本的比较，目前来看英伟达在推理领域依旧有着明显的优势，但云厂商可能会继续冒着支出更多的代价，坚定推进自研算力。利好云厂商的 ASIC 合作伙伴，以及其配套产业链：博通、amd、mrvl、arista、cls、中际旭创、新易盛、cohr、lite 等。

3、通信：deepseek 对通信的影响比较综合，通常训练集群通信要求相较于同等规模的推理集群更高，但考虑到性价比问题，目前众多训练集群也兼顾了推荐算法、推理、科研、金融交易等其他功能，我们认为训练集群依然将坚定的向着规模更大的方向发展。同时，推理需求需要考虑规模效应，预计云厂商主导的推理市场依旧将占主导地位，端侧目前算力有限，不构成太大影响，预计推理集群也将延续过去发展趋势。综合来看，无论是训练还是推理，依旧将使用到更多的高端交换机和光模块，我们倾向于推荐产品有优势、在手订单充足、产能扩展顺利的标的，建议关注：博通、arista、cls、中际旭创、新易盛、cohr、lite、ciena、德科立。

5. 英伟达、亚马逊、微软同日接入 DeepSeek

据大河网报道，北京时间 1 月 31 日，英伟达宣布 DeepSeek-R1 模型登陆 NVIDIANIM。同一时段内，亚马逊和微软也接入 DeepSeek-R1 模型。英伟达称，DeepSeek-R1 是最先进的大语言模型。

韩国 Mirae Asset Securities Research 的一名分析师在 X 撰写长文分析称：“这一突破是通过实施大量细粒度优化和使用英伟达的汇编式 PTX 编程，而非通过英伟达 CUDA 中的某些功能来实现的。”

也就是说 DeepSeek 在研发大模型时绕过了 CUDA。CUDA（Compute Unified Device Architecture，统一计算架构），是由英伟达开发的一种通用编程框架，它允许开发者利用英伟达的图形处理器（GPU，Graphics Processing Unit）进行通用计算。

在 DeepSeek-V3 的技术博文中，DeepSeek 表示其使用了英伟达的 PTX（Parallel Thread Execution）语言。

假如 DeepSeek 的开发者能够很好地使用 PTX（Parallel Thread Execution）语言，那么相比使用 CUDA 提供的编程接口，肯定可以更精细地控制 GPU 之间传输数据、权重和梯度等。但是，使用 PTX 写出来的代码非常复杂，且很难维护，因此需要专业度较高的开发者。

6. OpenAI 上线 o3-mini，首次向 ChatGPT 免费用户开放推理模型

当地时间 1 月 31 日，OpenAI 宣布正式推出推理模型 o3-mini，是其推理系列中最新、最具成本效益的模型，即日起可在 ChatGPT 和 API 中使用。

作为首款支持开发者高频需求功能的小型推理模型，OpenAI o3-mini 内置函数调用、结构化输出和开发者消息等专业功能，开箱即用，可直接投入生产环境。此外，开发者还可根据场景需求，灵活选择低、中、高三级推理强度，使模型在应对复杂挑战时能“深度思考”，或在需要快速响应时优先保证速度。

ChatGPT Plus、Team 及 Pro 用户即日可体验 o3-mini，企业用户将在一周后获得访问权限。即日起，免费版用户也可通过消息编辑器选择“推理”模式或重新生成回复来试用 o3-mini，这是 ChatGPT 首次向免费用户开放推理模型。

7. 风险提示

AI 发展不及预期，算力需求不及预期，市场竞争风险。

免责声明

国盛证券有限责任公司（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，但本公司及其研究人员对该等信息的准确性及完整性不作任何保证。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，可能会随时调整。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。

本公司力求报告内容客观、公正，但本报告所载的资料、工具、意见、信息及推测只提供给客户作参考之用，不构成任何投资、法律、会计或税务的最终操作建议，本公司不就报告中的内容对最终操作建议做出任何担保。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。投资者应当充分考虑自身特定状况，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

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。

分析师声明

本报告署名分析师在此声明：我们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，本报告所表述的任何观点均精准地反映了我们对标的证券和发行人的个人看法，结论不受任何第三方的授意或影响。我们所得报酬的任何部分无论是在过去、现在及将来均不会与本报告中的具体投资建议或观点有直接或间接联系。

投资评级说明

投资建议的评级标准		评级	说明
评级标准为报告发布日后的 6 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；新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以摩根士丹利中国指数为基准，美股市场以标普 500 指数或纳斯达克综合指数为基准。	股票评级	买入	相对同期基准指数涨幅在 15%以上
		增持	相对同期基准指数涨幅在 5%~15%之间
		持有	相对同期基准指数涨幅在-5%~+5%之间
		减持	相对同期基准指数跌幅在 5%以上
	行业评级	增持	相对同期基准指数涨幅在 10%以上
		中性	相对同期基准指数涨幅在-10%~+10%之间
		减持	相对同期基准指数跌幅在 10%以上

国盛证券研究所

北京
地址：北京市东城区永定门西滨河路 8 号院 7 楼中海地产广场东塔 7 层
邮编：100077
邮箱：gsresearch@gszq.com

上海
地址：上海市浦东新区南洋泾路 555 号陆家嘴金融街区 22 栋
邮编：200120
电话：021-38124100
邮箱：gsresearch@gszq.com

南昌
地址：南昌市红谷滩新区凤凰中大道 1115 号北京银行大厦
邮编：330038
传真：0791-86281485
邮箱：gsresearch@gszq.com

深圳
地址：深圳市福田区福华三路 100 号鼎和大厦 24 楼
邮编：518033
邮箱：gsresearch@gszq.com