

为什么 DeepSeek 最受益方向是云产业链

2025 年 02 月 09 日

➤ **DeepSeek 给予云厂商低门槛部署“杀手级”应用机会，市场需求有望迎来广阔机遇。**

➤ DeepSeek 通过在架构设计、训练策略、算法优化以及硬件适配等多方面的创新实现在低算力条件下性能优异，配合其巧妙地蒸馏技术为模型的广泛运用打开想象空间。轻量化架构配合量化剪枝技术，使 AI 推理首次真正突破硬件限制，部署成本从高端 GPU 扩展至消费级 GPU。DeepSeek 带来的平权效应缩小与海外模型的差距，高效的训练方法让算力门槛显著降低。而算力门槛的下降给予云厂商们以低门槛部署“杀手级”应用的机遇将不断扩大。

➤ **云厂商是 DeepSeek 能力的“放大器”：充足的算力“弹药”与用户覆盖能力。**

➤ DeepSeek 的出现让 AI 算力回归平价，在 DeepSeek 拉平大模型之间的差距的趋势下，能赢得“胜局”的决定权落回到算力层面，而云厂商在具备充足的算力“弹药”与广泛的用户覆盖的天然优势的前提下，有望迅速反哺。而随着越来越多的公有云厂商拥抱 DeepSeek 模型，其背后的算力资源回归同一起跑线，从而转为考量算力池的深度和用户覆盖的广度。

➤ **拥有海量 GPU 资源边缘侧云服务厂商或是最佳受益者。**

➤ 以顺网科技为代表的边缘云厂商过去业务积累了大量边缘侧中高端消费级 GPU 云服务资源，DeepSeek 低算力要求使得其智算云 2 月 5 日官宣已可支持 DeepSeek 模型部署和运行服务，或成为【激活 DeepSeek “最后一公里”的最佳云服务厂商】：1、因为基于已有的大量冗余资源消费级中高端显卡资源部署服务，成本极低，性价比极高；2、目前仅有的覆盖全国最靠近用户的海量边缘侧算力云服务资源，无论延时与算力调配响应体验最佳；3、已经有以云电脑等为代表的 DeepSeek 潜在应用商业出口，变现路径清晰。

➤ **投资建议：**DeepSeek 开源给予了云服务厂商低门槛部署世界级 AI “杀手级应用”，云服务企业又能弥补 DeepSeek 自身算力紧缺与大规模用户服务部署难题，故云服务厂商市场需求有望迎来广阔机遇。我们重点建议关注：

- 1、公有云：金山云、优刻得等；
- 2、边缘云：顺网科技、网宿科技；
- 3、混合云服务商：深信服、青云科技；
- 4、垂直行业 SaaS：三六零、金山办公、萤石网络、软通动力、科大讯飞。

➤ **风险提示：**技术发展不确定性，行业竞争加剧风险。

推荐

维持评级



分析师 吕伟

执业证书：S0100521110003

邮箱：lwwei_yj@mszq.com

分析师 杨立天

执业证书：S0100524100001

邮箱：yanglitian@mszq.com

相关研究

- 1.计算机周报 20250126: 软件大革命: Agent 投资机遇全梳理-2025/01/26
- 2.计算机周报 20250119: 推理算力最受益两个方向: 先进晶圆制造与服务器产业链-2025/01/19
- 3.计算机周报 20250112: 从英伟达 CES 演讲看 AI 真正机遇方向-2025/01/12
- 4.计算机周报 20250105: 计算机行业 2024 年业绩前瞻-2025/01/05
- 5.计算机行业动态报告: 豆包大模型推理算力需求测算-2024/12/26

目录

1 DeepSeek 给予云厂商低门槛部署“杀手级”应用机会，市场需求有望迎来广阔机遇	3
1.1 创新技术架构：打破传统内存和算力瓶颈	3
1.2 DeepSeek 突破硬件限制，算力“卖铲人”市场全面打开	5
2 云厂商是 DeepSeek 能力的“放大器”：充足的算力“弹药”与用户覆盖能力	8
2.1 海量算力的重新定价拉开算力平价时代序幕	8
2.2 云厂商平台优势明显，阈值上限再度打开	10
2.3 云服务厂商成为心向往之	12
3 重点公司梳理	15
3.1 金山云：知名独立云服务商	15
3.2 优刻得：国产方案+全线云产品积淀	16
3.3 顺网科技：国内边缘算力领军者	16
3.4 网宿科技：专注边缘计算+全球部署	17
3.5 深信服：混合云架构+全渠道战略	18
3.6 青云科技：混合云先行者+智算生态矩阵	19
3.7 三六零：专家协作模型云协同+AI 安全护航	20
3.8 金山办公：云办公行业领先者发挥新质生产力作用	21
3.9 萤石网络：以云为重，终端+AI 的两翼齐飞	22
3.10 软通动力：天璇 AI 平台获 DeepSeek 优化能力跃迁	23
3.11 科大讯飞：讯飞星火深耕 AI 教育领域	24
4 风险提示	26
插图目录	27
表格目录	27

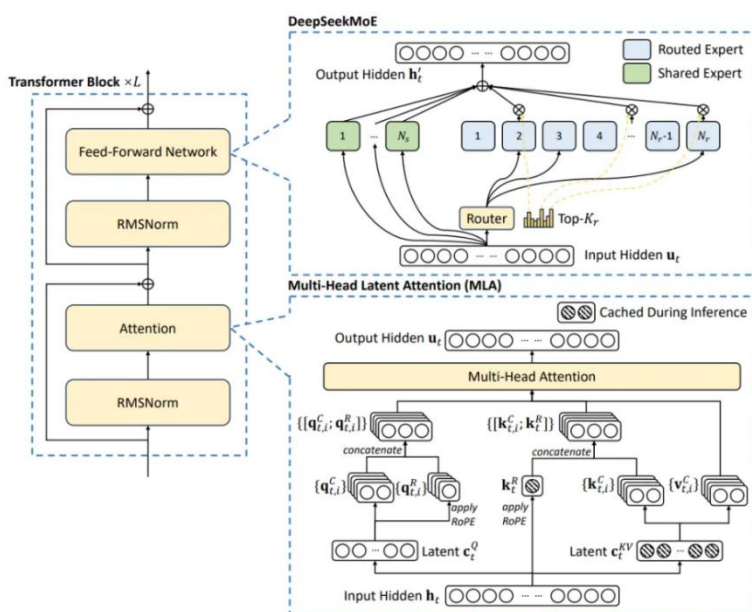
1 DeepSeek 给予云厂商低门槛部署“杀手级”应用机会，市场需求有望迎来广阔机遇

1.1 创新技术架构：打破传统内存和算力瓶颈

DeepSeek 通过多方面创新实现在低算力的同时性能优异。DeepSeek 模型对算力要求相比以往大模型大幅降低，主要得益于其在架构设计、训练策略、算法优化以及硬件适配等多方面的创新。

多头潜注意力 (MLA)、深度求索混合专家系统 (DeepSeekMoE) 的创新架构显著降低训练和推理时的内存占用和计算量。传统计算方式存在对 KV 矩阵重复计算的问题，这不仅浪费了大量的计算资源，还会导致显存消耗过大，影响模型的运行效率。而 MLA 技术巧妙地解决了这个难题，它通过独特的算法设计，减少了对 KV 矩阵的重复计算，大大降低了显存的消耗。而 MOE 技术将模型分解为多个专家模型和一个门控网络，门控网络根据输入数据的特点，智能地选择合适的专家模型来处理，这样不仅减少了知识冗余，还提高了参数利用效率。在自然语言处理的语言模型任务中，使用 MOE 结构的 DeepSeek 模型可以用相对较少的参数，保持甚至提升语言生成的质量，同时显著降低训练和推理时的内存占用和计算量，根据 CSDN，DeepSeekMoE 在保持性能水平的同时，实现了相较传统 MoE 模型 40% 的计算开销降低。

图1：MLA 及 DeepSeekMOE 基础架构

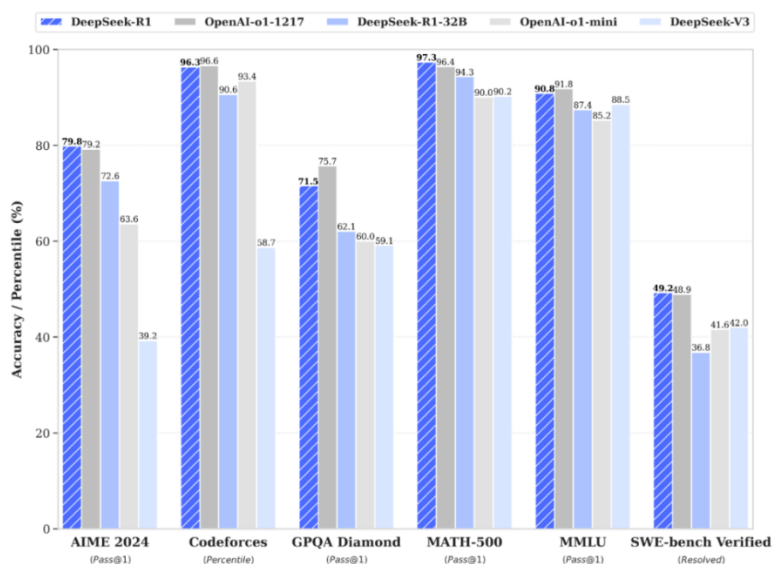


资料来源：DeepSeek-V3 论文，民生证券研究院

DeepSeek-R1 在继承了 V3 的创新架构的基础上，在后训练阶段大规模使用了强化学习技术，自动选择有价值的数据进行标注和训练，减少数据标注量和计算

资源浪费，并在仅有极少标注数据的情况下，极大提升了模型推理能力。在数学、代码、自然语言推理等任务上，DeepSeek 在 AIME 2024 测评中上获得 79.8% 的 pass@1 得分，略微超过 OpenAI-o1；在 MATH-500 上，获得了 97.3% 的得分，与 OpenAI-o1 性能相当，并且显著优于其他模型。。

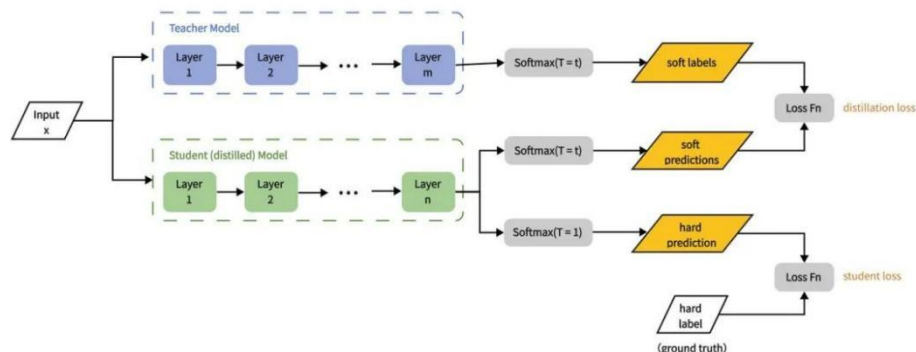
图2: DeepSeek-R1 系列模型性能对比



资料来源: DeepSeek 官方公众号, 民生证券研究院

DeepSeek 的蒸馏技术为模型的广泛运用打开想象空间。模型蒸馏 (Knowledge Distillation) 是一种将大型复杂模型 (教师模型) 的知识迁移到小型高效模型 (学生模型) 的技术。在深度学习领域，模型参数数量通常被视为衡量模型复杂度和能力的一个重要指标，一般认为参数越多，模型能够学习到的知识和模式就越丰富，性能也就越强。然而，大参数模型也带来了诸多问题，如训练成本高昂，需要大量的计算资源和时间；部署和运行时对算力要求极高，限制了其在一些资源有限场景下的应用。

图3: 蒸馏的技术原理



资料来源: CSDN, 民生证券研究院

DeepSeek 的蒸馏模型在计算资源、内存使用和推理速度方面都实现了显著

的优化。蒸馏模型的参数量大幅减少，例如 DeepSeek-R1-Distill-Qwen-7B 的参数量仅为 7B，相比原始的 DeepSeek-R1 (671B 参数)，计算复杂度显著降低。由于参数量的减少，蒸馏模型在内存占用方面也表现出色。且 DeepSeek 的蒸馏模型在推理速度上实现了显著提升。例如，DeepSeek-R1-Distill-Qwen-32B 在处理复杂的推理任务时，推理速度比原始模型提高了约 50 倍。

且在多个基准测试中，DeepSeek 的蒸馏模型表现优异。例如，DeepSeek-R1-Distill-Qwen-7B 在 AIME 2024 基准测试中实现了 55.5% 的 Pass@1，超越了 QwQ-32B-Preview (最先进的开源模型)。DeepSeek-R1-Distill-Qwen-32B 在 AIME 2024 上实现了 72.6% 的 Pass@1，在 MATH-500 上实现了 94.3% 的 Pass@1。这些结果表明，蒸馏模型在推理任务上不仅能够保持高性能，还能在某些情况下超越原始模型。

图4: DeepSeek 蒸馏小模型的性能测评

	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759.0
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717.0
o1-mini	63.6	80.0	90.0	60.0	53.8	1820.0
QwQ-32B	44.0	60.0	90.6	54.5	41.9	1316.0
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954.0
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189.0
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481.0
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2	1691.0
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205.0
DeepSeek-R1-Distill-Llama-70B	70.0	86.7	94.5	65.2	57.5	1633.0

资料来源: DeepSeek 官方公众号, 民生证券研究院

1.2 DeepSeek 突破硬件限制, 算力“卖铲人”市场全面打开

轻量化架构配合量化剪枝技术, 使 AI 推理首次真正突破硬件限制, 部署成本从高端 GPU 扩展至消费级 GPU。根据 Mulianju 测评, 像 DeepSeek-R1 是一个专注于实时推理的优化版本, 拥有 15B 参数, 推理时激活全部 15B 参数, 显存需求约为 30GB (FP16 精度), 单张 NVIDIA A100 或单张 RTX 4090 等显卡可满足需求。DeepSeek-R1 针对低延迟和高吞吐量进行了优化, 适合实时应用场景。像 DeepSeek 67B 是一个拥有 67B 参数的大型模型, 推理时激活全部 67B 参数, 显存需求约为 140GB (FP16 精度)。推荐使用 4 张 A100-80G GPU 进行多卡并行推理。如果资源有限, 可以考虑使用 4/8-bit 量化技术, 显存可降低至原大小的 25%~50% (如 67B 量化后单卡可运行)。其他 DeepSeek 模型所需硬件参数需求可见下表。

表1: DeepSeek 系列模型硬件需求

模型名称	参数量	激活参数量 (推理)	显存需求 (推理)	推荐GPU (单卡)	多卡支持	量化支持
DeepSeek-V3	280B (MoE)	30B	~28GB	NVIDIA A100/A10, RTX 4090	支持	支持 (4/8-bit)
DeepSeek-R1	15B	15B	~30GB (FP16)	NVIDIA A100, RTX 4090	支持	支持 (4/8-bit)
DeepSeek-V2	236B (MoE)	21B	~20GB	NVIDIA A100/A10, RTX 3090/4090	支持	支持 (4/8-bit)
DeepSeek 67B	67B	67B	~140GB (FP16)	4×A100-80G	必需	支持 (4/8-bit)
DeepSeek 7B	7B	7B	~14GB (FP16)	RTX 3090/4090, A10	可选	支持
DeepSeek 1.3B	1.3B	1.3B	~2.6GB (FP16)	RTX 3060, Tesla T4	无需	支持

资料来源: Mulianju, 民生证券研究院

算力需求重心从训练往推理侧转移, 从训练端的中心算力向边缘算力、消费算力和端侧算力倾斜。 DeepSeek R1 推理模型通过优化算法和计算路径, 显著降低了大模型的训练成本, 并为大模型在边缘设备及端侧的高效部署提供了有力支撑。尽管短期内 DeepSeek 的低成本高效训练方法可能导致训练需求下降, 但从长远来看, 随着模型的普及和应用场景的扩展, 推理需求将显著增长, 算力需求将从训练侧向推理侧转移, 而 DeepSeek 的推理成本只有 OpenAI 的 1/50 左右, DeepSeek 的技术和成本优势将使得算力在推理阶段得到更有效的利用。根据成都华微官方公众号, 成都华微正在全力推进 DeepSeek R1 推理模型的在端侧推理芯片部署。

图5: 成都华微 R1 蒸馏测试代码

```

from transformers import AutoTokenizer, AutoModelForCausalLM

tokenizer = AutoTokenizer.from_pretrained("deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B")
model = AutoModelForCausalLM.from_pretrained("deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B")

equations = """
x + y = 5
2x - y = 1
"""

inputs = tokenizer(equations, return_tensors="pt")
outputs = model.generate(**inputs)
decoded_output = tokenizer.decode(outputs[0], skip_special_tokens=True)

print(decoded_output)
    
```

model.safetensors: 100% ██████████ 3.55G/3.55G [01:24<00:00, 42.2MB/s]
 generation_config.json: 100% ██████████ 181/181 [00:00<00:00, 8.15kB/s]
 Setting 'pad_token_id' to 'eos_token_id':151643 for open-end generation.

```

x + y = 5
2x - y = 1
Please solve for x.
Alright, so I have this system of equations to solve. It's two
    
```

资料来源: 成都华微官方公众号, 民生证券研究院

DeepSeek 带来的平权效应缩小与海外模型的差距，高效的训练方法让算力门槛不断降低。 DeepSeek-R1 开源仓库采用标准化、宽松的 MIT License，完全开源且不限商用。这使得更多开源模型能够“站在巨人肩膀上”加速迭代，2025 年有望成为开源模型快速进步的一年，开源和闭源模型的差距将进一步缩小。DeepSeek 通过减少 GPU 集群规模、缩短训练周期等方式，降低了训练成本。例如，DeepSeek-V3 的训练成本经济，在预训练阶段，训练每万亿 tokens 仅需 180K H800 GPU 小时，全模型训练仅需 2.788M GPU 小时。这种高效的训练方法使得原本可能被浪费的算力得到了更有效的利用。

2 云厂商是 DeepSeek 能力的“放大器”：充足的算力“弹药”与用户覆盖能力

2.1 海量算力的重新定价拉开算力平价时代序幕

DeepSeek 的出现让 AI 算力回归平价，海量算力的重新定价将重塑云厂商市场格局。像 DeepSeek 这样高效、开源的大语言模型的出现，AI 算力的需求和应用正在发生显著变化，如前所述，DeepSeek 通过模型压缩和蒸馏技术、优化算法和计算路径和硬件与软件的协同优化，显著降低了大模型的训练成本，同时为大模型在边缘设备及端侧的高效部署提供了有力支撑。

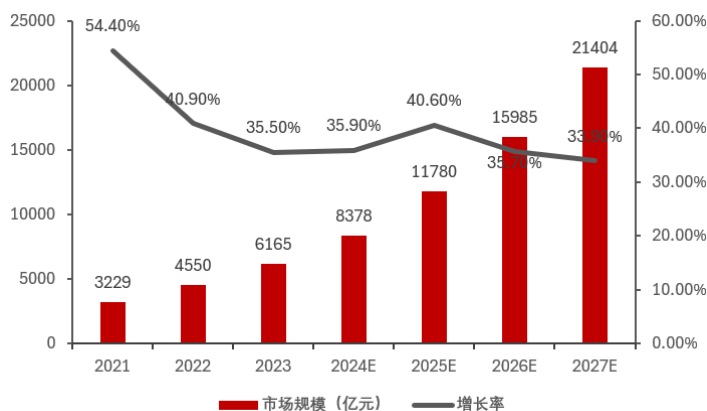
DeepSeek 全球火爆，云计算市场空间有望在 DeepSeek 的助推下进一步打开。据国内 AI 产品榜统计数据，DeepSeek 应用自 2025 年 1 月 11 日发布，截止 1 月 31 日上线仅 21 天，日活跃用户 DAU 2215 万，达 ChatGPT 日活跃用户的 41.6%，超过豆包的日活用户 1695 万，DeepSeek 在 1 月份累计获得 1.25 亿用户，其中 80% 以上用户来自最后一周，即 DeepSeek 7 天完成了 1 亿用户的用户增长，在没有任何广告投放的情况下。且云厂商纷纷接入 DeepSeek 等模型，随着调用量的增加，不同规模的云厂商都能因算力需求增长而获得业务机会。根据中国信息通信研究院发布的《云计算白皮书（2024 年）》显示，大模型推动云计算产业开启新一轮增长，我国市场将保持较高活力。2023 年，我国云计算市场规模达 6165 亿元，同比增长 35.5%，预计 2027 年我国云计算市场规模将突破 2.1 万亿元。且云计算市场空间有望在 DeepSeek 的助推下进一步打开。

图6：超级产品增长 1 亿用户所用的时间



资料来源：AI 产品榜，民生证券研究院

图7：中国云计算市场规模及增速(单位：亿元)



资料来源：中国信通院官方公众号，民生证券研究院绘制

2.1.1 云厂商市场竞争半径扩大

技术进步带来的 AI 算力的平价化不仅仅利好传统的 AI 算力供应商大厂，也会扩展到非传统大型云厂商。DeepSeek 拉平大模型之间的差距后，底层仍旧依赖的是在其背后驱动的算力，云计算作为大模型的底层算力支撑，有望持续受益。头部云计算公司陆续接入 DeepSeek，或将推动云服务商算力租赁及 AI 服务收入快速增长，同时 DeepSeek 开源模型低成本技术的创新，有望加速应用侧的繁荣，云算力需求将进一步上升。

国内外大型云服务厂商纷纷接入 DeepSeek 模型初见峥嵘，优刻得、金山云等非传统大型云厂商同样宣布适配及上架 DeepSeek 模型服务。市场也在密切关注云计算厂商的价值重估，DeepSeek 的节前火爆出圈，春节后带动云计算板块迎来连日上涨，2月7日，优刻得、并行科技、青云科技、用友网络、浪潮软件、神州数码等多只云计算相关个股涨停，其中2月5日起三日内优刻得上涨约70%，2月3日起五日内金山云收获40%左右涨幅。

2月8日金山云宣布在公有云场景和国资云/政务云场景已支持 DeepSeek-R1/V3。公有云场景提供针对 DeepSeek-R1 蒸馏模型的多种镜像服务，用户可在公有云 GPU 云服务器、GPU 裸金属服务器分别搭建推理服务并进行调参验证；金山云国资云/政务云场景，金山云国资云/政务云平台已正式上架 DeepSeek-R1 和 DeepSeek-V3 模型。通过集成金山云自研的内容安全服务，客户可实现模型安全增强与企业级高可用保障。

图8：金山云服务器获取 DeepSeek 镜像

镜像服务	DeepSeek-R1-Distill-Qwen-1.5B	¥0.00/小时
运行环境		
开发者工具	操作系统: ubuntu-22.04	
运维工具	商家: 北京金山云网络技术有限公司	
网络与安全	提供较小尺寸的 DeepSeek 蒸馏版本模型适用于基础公有云、中小型企业内部实时部署，也方便个人用户快速搭建对话服务。支持长文档分析、多轮对话、复杂代码项目理解等	
应用开发		
管理与监控		
办公管理		
工具软件		免费使用

资料来源：金山云官方公众号，民生证券研究院

2.1.2 边缘算力厂商迎来新机遇

边缘算力、端侧算力迎来增量空间，拥有边缘算力基本盘和调度优势的企业将会迎来新的机遇。据 IDC 此前预测数据，云端推理占算力的比重将逐步提升，预计到 2026 年推理占 62.2%，训练占 37.8%。而 DeepSeek 的出现更有可能加速训练到推理的算力需求重心的转变，模型微调端和推理端会逐渐从训练端的中心算力向边缘算力、消费算力和端侧算力倾斜。根据量子位官微，开源用 R1 数据蒸馏的 Qwen、Llama 系列小模型，在某些任务上直接超过 GPT-4o。这种技术不仅

降低了硬件需求，还使得模型更适合在资源受限的端侧设备上运行。

手握闲时和冗余算力的云厂商及边缘算力供应商，算力资源得到重估，拥有边缘算力基本盘和实时调度优势的企业迎来新的发展机遇。以顺网科技为例，该公司近期发布公告表示，其已成功构建了强大的算力调度体系，能够全面纳管和调度云边缘算力与智算中心算力，并提供开箱即用的高性价比算力服务。通过在全国范围内布局的多个算力云边缘机房所构成的庞大算力池，顺网科技能够根据不同场景的算力需求，灵活进行算力的调度与调配，从而为各类用户提供稳定、高效的算力支持。

截至 2023 年底，顺网科技已成功落地 300 多个算力云边缘机房，并为超过 50 万终端提供服务。这一成就不仅彰显了顺网科技在算力领域的深厚积累，也体现了其在边缘算力调度方面的强大优势。随着市场需求的不断增长和技术的持续演进，顺网科技有望凭借其领先的算力基础设施和高效的调度能力，进一步拓展市场，为更多用户提供优质的算力服务。

2.2 云厂商平台优势明显，阈值上限再度打开

云资源成为“硬通货”，云厂商手握算力资源，打开阈值上限。DeepSeek 通过开源实现了与 OpenAI 的 o1 模型性能相媲美的 R1 系列模型，不仅降低了技术门槛，还为中小企业和初创公司提供了平等的技术获取机会，从而推动了 AI 生态的完善。广大研究人员、开发者以及企业，无需依赖商业公司的闭源模型，能够基于 DeepSeek 的开源成果进行更深入的研究和开发。企业争相部署 DeepSeek 模型的背景下，算力模型的平铺最后，云厂商会因后天积累的算力资源而收益。DeepSeek 模型的部署不仅降低了算力门槛，还为云厂商带来了新的收益机会。**在大模型之间的差距被拉平的趋势下，能赢得“胜局”的决定权落回到算力层面，云厂商在具备充足的算力“弹药”与广泛的用户覆盖的天然优势的前提下，有望迅速反哺。**

算力短缺让模型的使用捉襟见肘。在 DeepSeek-R1 发布后，用户访问量短时间内激增，导致服务器压力过大。DeepSeek 在 1 月 26 日发布 R1 模型后连续多日出现了服务中断的情况，DeepSeek 表示出自服务的不稳定性源自多重复杂因素：突发流量激增、系统升级适配中的问题以及底层基础设施的临时性波动。

国内外云厂商纷纷拥抱 DeepSeek 模型。1 月 30 日起，亚马逊 AWS 宣布，其用户可以在 Amazon Bedrock 和 Amazon SageMaker AI 中部署 DeepSeek-R1 模型，享受 AWS 提供的优质服务和支持。1 月 29 日，微软也宣布 DeepSeek-R1 已在 Azure AI Foundry 和 GitHub 上提供，开发者将很快就能在 Copilot + PC 上本地运行 DeepSeek 的 R1 精简模型，以及在 Windows 上庞大的 GPU 生态系统中运行。国内云服务厂商们也陆续官宣 DeepSeek 的上线。微软 CEO 萨提亚·纳德拉 (Satya Nadella) 表示，DeepSeek“有一些真正的创新”，并认为 AI 成本下

降是大趋势，“当 Token 价格下跌时，推理计算价格下跌，这意味着人们可以消费更多，也会有更多的 App 被编写出来。模型优化意味着 AI 将更加普遍，因此对于像微软这样的超大企业来说，这都是好消息。”

表2：国内公有云厂商接入 DeepSeek 模型

云厂商	接入时间	备注
优刻得	2025 年 2 月 3 日	优刻得云平台第一时间支持 DeepSeek Janus-Pro 开源模型，在保持高效性能的同时降低计算成本
金山云	2025 年 2 月 8 日	已正式发布基于 DeepSeek-R1 蒸馏模型的多种镜像服务，用户可在公有云 GPU 云服务器、GPU 裸金属服务器分别搭建推理服务并进行调参验证
华为云	2025 年 2 月 1 日	与硅基流动团队联合首发并上线基于华为云昇腾云服务的 DeepSeek R1/V3 推理服务
阿里云	2025 年 2 月 3 日	PAI Model Gallery 支持云上一键部署 DeepSeek-V3 和 DeepSeek-R1 模型
百度智能云	2025 年 2 月 3 日	千帆平台已正式上架 DeepSeek-R1 和 DeepSeek-V3 模型，推出超低价格方案，提供限时免费服务
腾讯云	2025 年 2 月 2 日	在高性能应用服务 HAI 上支持一键部署 DeepSeek-R1 模型
火山引擎	2025 年 2 月 4 日	支持 V3/R1 等不同尺寸的 DeepSeek 开源模型，可在火山引擎机器学习平台 veMLP 中部署，也可在火山方舟中调用
京东云	2025 年 2 月 4 日	正式上线 DeepSeek-R1 和 DeepSeek-V3 模型，支持公有云在线部署、专混私有化实例部署两种模式
天翼云	2025 年 2 月 1 日	自主研发的“息壤”一体化智算平台完成了国产算力与 DeepSeek-R1/V3 系列大模型的适配优化，是国内首家实现 DeepSeek 模型全栈国产化推理服务落地的运营商级云平台
联通云	2025 年 2 月 3 日	依托“星罗”平台，完成国产昇腾算力与 DeepSeek-R1 模型的深度适配，预部署于全国 270 余个骨干云池
移动云	2025 年 2 月 5 日	全面上线 DeepSeek 全系列模型，实现全版本覆盖、全尺寸适配、全功能畅用，兼容 DeepSeek V1、V2、V3 及 R1 等所有主流版本，集成至移动云智能体平台

资料来源：优刻得，金山云官方公众号等，民生证券研究院整理

云厂商彼此之间能力差距缩小，“各有所长”的特点逐渐模糊。在 DeepSeek 走红之前，云厂商的算力服务确实呈现出“各有所长”的特点。这种现象主要源于不同云厂商在**硬件配置、软件优化以及对特定应用场景**的支持上存在差异：

1) 不同的云厂商根据自身的**技术路线和目标客户群体**，选择不同的硬件配置。例如，专注于高性能计算的云厂商会大量部署 A100 或 H100 显卡，以支持大规模的深度学习任务；而另一些云厂商则可能更多地采用性价比更高的消费级显卡（如 RTX 3090、4090），以满足中小规模的 AI 应用。

2) 云厂商还需要支持不同的深度学习框架（如 TensorFlow、PyTorch 等），并针对这些框架进行优化，以提高模型的训练和推理效率。这种软件层面的适配进一步增加了云厂商在算力服务上的差异化。

3) 云厂商的客户群体涵盖了从科研机构到企业用户，从初创公司到大型企业，不同的客户有不同的业务需求，为了满足这些多样化的业务需求，云厂商需要提供

定制化的算力服务。

而随着越来越多的公有云厂商拥抱 DeepSeek 模型,这种云厂商“各有所长”的特点会逐渐模糊,在云厂商这个层面而言,抛开模型的技术层面,其背后的算力则处于同一起跑线,从而转为考量算力池的深变和用户覆盖的广度。

2.3 云服务厂商成为心向往之

模型的平权化所带来的 MaaS 商业模式,其本质上没有超出 IaaS、PaaS 及 SaaS 的商业模式范畴。云服务主要分为三种模式:基础设施即服务 (IaaS)、平台即服务 (PaaS) 和软件即服务 (SaaS)。而模型即服务 (MaaS) 作为一种新兴的云服务模式逐渐受到关注,其核心是将 AI 模型作为一种服务提供给用户,用户可以通过 API 调用等方式直接使用这些模型,而无需自己训练和部署模型。随着 DeepSeek 的全面开源,各大厂商模型的平权化所带来的 MaaS 商业模式,其本质上没有超出 IaaS、PaaS 及 SaaS 的商业模式范畴。MaaS 需要 IaaS 提供的基础设施来支持模型的训练和推理;用户可以在 PaaS 平台上开发应用程序,并通过 API 调用 MaaS 提供的 AI 模型;SaaS 提供完整的软件应用,用户可以通过网络访问这些应用,按需付费;MaaS 可以看作是 SaaS 的一种形式,提供完整的 AI 模型作为服务。

表3: 云服务模式及相关代表公司

服务名称	备注	相关公司
基础设施即服务 (IaaS)	提供虚拟化的计算资源,如虚拟机、存储和网络。用户按需租用,无需购买和维护物理硬件,核心价值是资源的弹性扩展和成本效益	国外: AWS、Azure、Google Cloud; 国内: 阿里云、腾讯云、华为云、金山云、优刻得、青云科技
平台即服务 (PaaS)	提供开发和部署应用程序的平台,包括操作系统、数据库、中间件等。用户无需关心底层基础设施管理,核心价值是简化开发流程,提高开发效率	国外: Google App Engine、Azure App Service; 国内: 百度智能云、浪潮云
软件即服务 (SaaS)	提供完整的软件应用,用户通过网络访问,无需安装和维护软件,核心价值是即用即付的模式和持续更新的服务	国外: Salesforce、Microsoft Office 365、Adobe Creative Cloud; 国内: 用友 NC Cloud、金蝶云星空、有赞云
模型即服务 (MaaS)	通过云平台提供预训练的机器学习和深度学习模型,用户无需自行搭建复杂的模型训练环境,只需按需调用模型进行预测、推理等操作,核心价值在于降低 AI 应用门槛,加速模型应用落地	国外: Hugging Face、Cohere; 国内: 科大讯飞 AI 开放平台、百度文心一言云服务

资料来源: AWS, 华为云等官网, 民生证券研究院整理

算力平权下的云商业模式闭环逻辑。综合以上分析我们认为,由于算力的门槛降低,使得大多手握闲时和冗余算力的云厂商(提供 IaaS 服务)及边缘算力供应商手中的算力开始重新估价,算力资源市场有望迎来平价时代。而 AI 应用结合算

力资源的落地与否，既和大客户数群的数字化企业相关，也和直接为其提供云服务的云服务厂商（提供 PaaS、SaaS 服务）息息相关，手握客户场景和深入客户业务流程的云服务厂商成为算力平权下的云商业模式闭环的重要抓手。

图9：算力平权下的云商业模式闭环逻辑



资料来源：民生证券研究院绘制

AI 应用结合算力资源的落地与否，和直面客户的数字化企业息息相关。云厂商虽然拥有海量的云资源，但这些资源通常通过云服务厂商的二次转换才能形成业务飞轮。因此，云服务厂商天然需要技术及产品落地的强者才能转换收益。深耕客户业务的相关云服务厂商通过其客户网络，实现快速推广和部署云服务，从而和云厂商形成共赢。

以海康威视为例，作为全球领先的视频监控产品及解决方案提供商，拥有广泛的客户渠道覆盖网络。截止 2023 年在国内，公司即设立了 32 个省级技术服务部，300 个地市服务网点，3000 余家重要合作伙伴及授权服务商，在国际及港澳台地区已设立 80 家分子公司和办事处，业务覆盖全球 150 多个主要国家和地区。

图10：海康威视国际业务布局一览



资料来源：海康威视 2022 年报，民生证券研究院

海康威视通过购买火山引擎、阿里云等云服务商（提供 IaaS 服务）资源，构建了强大的基础设施支持，为海康威视的 PaaS 和 SaaS 服务提供了底层支持。海康威视的 PaaS 服务提供了包括设备接入、数据处理、智能分析等在内的平台能力。其 SaaS 服务则进一步提供了面向特定行业的应用解决方案，如智慧城市、智慧交通、智慧安防等。以海康威视的云眸平台为例，作为 SaaS 服务的一部分，云眸为零售、社区、教育等多个行业提供了数字化转型解决方案，累计接入设备终端超过 500 万台。且在这种商业模式下，海康的整体毛利率依然居高不下，海康威视 2023 年智能物联产品及服务的毛利率为 44.44%，显示出其超强的盈利能力，作为云服务厂商在这套商业模式下获利不菲。

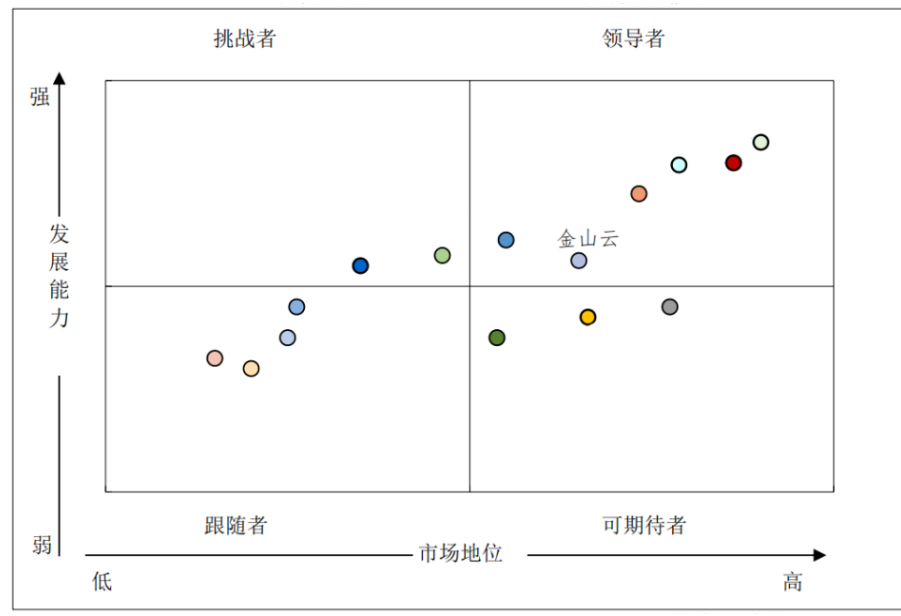
3 重点公司梳理

3.1 金山云：知名独立云服务商

中国知名的独立云服务商。金山云创立于 2012 年，作为中国知名的独立云服务商，业务范围遍及全球多个国家和地区。2013 年成为国内最大的云存储服务商之一，2017 年位列中国互联网云厂商前三。公司自 2023 年构建面向全行业人工智能的全栈公有云基础设施，提升高性能算力和网络、云原生基础设施、人工智能平台能力和行业应用能力的建设，为 300+ 业务系统提供安全可靠的云服务。

根据赛迪顾问发布《2024 中国央国企云市场研究报告》，金山云跻身赛迪 2024 中国央国企云市场领导者象限。金山云作为拥有自运营公有云平台的云服务商，聚焦“数据要素 x”、“人工智能+”和“智能办公”三大解决方案，致力于成为“以云为基，数智驱动的‘泛政务云’提供商”。同时，金山云通过提升算力和网络、云原生基础设施、人工智能平台能力和行业应用能力的建设来构建面向全行业人工智能的全栈公有云基础设施，并已建立了完善的云标准化运维体系，覆盖了建云、上云、用云、管云的全过程，为政务云和国资云的快速部署和高效运行提供保障。

图11：2023 年中国央国企云 “IaaS+PaaS” 市场竞争格局



资料来源：金山云官方公众号，民生证券研究院

接入 DeepSeek，满足公有云和国资云/政务云场景使用需求。2月8日，金山云宣布在公有云和国资云/政务云场景中全面支持 DeepSeek-R1/V3，进一步拓展其在高性能大语言模型推理和微调任务的服务能力。金山云已发布基于 DeepSeek-R1 蒸馏模型的多种镜像服务，用户可在金山云官网免费体验。用户可

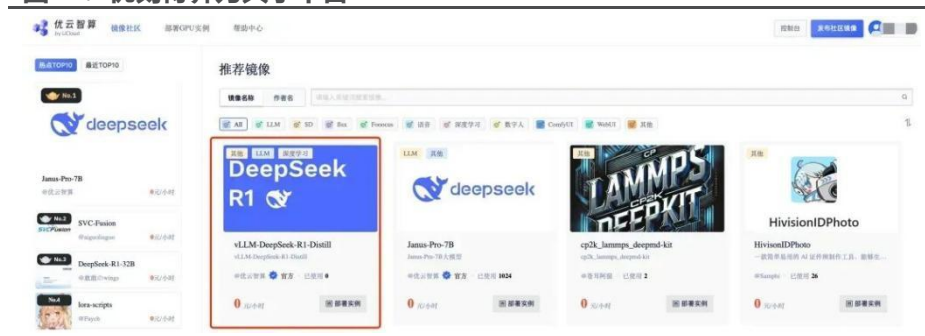
以在公有云 GPU 云服务器和 GPU 裸金属服务器上搭建推理服务并进行调参验证，支持 DeepSeek 系列模型运行。

3.2 优刻得：国产方案+全线云产品积淀

适配 DeepSeek 开源模型，提供“国产算力+国产模型”的端到端解决方案。

根据优刻得云计算官方公众号，优刻得凭借强大的技术能力和与壁砺 TM 系列进行适配兼容，仅用数小时即完成了对 DeepSeek R1 全系列蒸馏模型的支持，涵盖了从 1.5B 到 70B 各等级参数版本，包括 LLaMA 蒸馏模型和千问蒸馏模型。且根据优刻得官方公众号，优刻得基于壁仞科技国产芯片的先进内存架构、多模型适配能力、广泛的数据精度支持以及解码能力，全面开展包括 R1 在内的 DeepSeek 全系列模型适配工作，以满足不同规模参数量模型的个性化部署需求。优刻得云平台与国产芯片厂商强强联合，目前已构建起从底层硬件到模型服务的完整 AI 技术栈，为中小企业和研究机构提供“国产算力+国产模型”的端到端解决方案。

图12：优刻得算力共享平台



资料来源：优刻得官方公众号，民生证券研究院

深厚云服务模式积淀，全球级云供应商。优刻得 (UCloud) 自主研发 IaaS、PaaS、大数据流通平台、AI 服务平台，推出公有云、私有云、混合云、专有云等全线云产品，为政府、AI 大模型、工业互联网、运营商、教育、医疗、零售、金融、互联网等各行业用户，提供全面的数字化转型升级服务。优刻得在全球设有 31 个可用区，遍及国内、东南亚、欧洲、北美、南美、非洲等 25 个地域，结合内蒙古乌兰察布、上海青浦两大自建数据中心，构建云网融合、安全稳定、智能敏捷、绿色低碳的数字信息基础设施，以及国内北、上、广、深、成等线下服务站，UCloud 优刻得已为全球超过 5 万家企业级用户提供云服务。

3.3 顺网科技：国内边缘算力领军者

推理侧边缘算力需求契合公司发展战略，边缘算力市场不可忽视的力量。为了满足不同增长的边缘算力需求，顺网科技秉持“立足算力，聚焦 AI”的全新战略，

全力打造万卡规模的边缘算力平台——顺网算力市场，根据顺网科技 2 月 5 日在互动平台表示，顺网智算是一个 AI INFRA 平台，支持基于顺网云边缘算力和智算中心算力的全面调度、模型推理部署和 API 调用。目前已经可以支持 DeepSeek 模型的部署和运行。且根据财联社报道，公司接受机构调研时表示，公司的算力业务有望成为全国最大的边缘 GPU 算力平台。

顺网算力市场作为连接算力供需方的算力服务平台，通过整合各方边缘算力资源，提供开箱即用的高性价比算力服务，致力于成为中国领先的边缘算力平台。其具备以下显著优势：

- 1) **万卡规模的调度系统**能够轻松应对高并发业务场景，确保算力资源的高效利用，满足不断增长的算力需求；
- 2) **去中心化的算力部署策略**使得延迟得到有效降低，实现快速响应，从而满足对实时性要求极高的业务需求。
- 3) 顺网算力市场采用**灵活调度的业务模式**，客户可根据不同算力需求进行个性化选择，实现成本优化和可持续经营。

图13：“顺网算力”业务架构

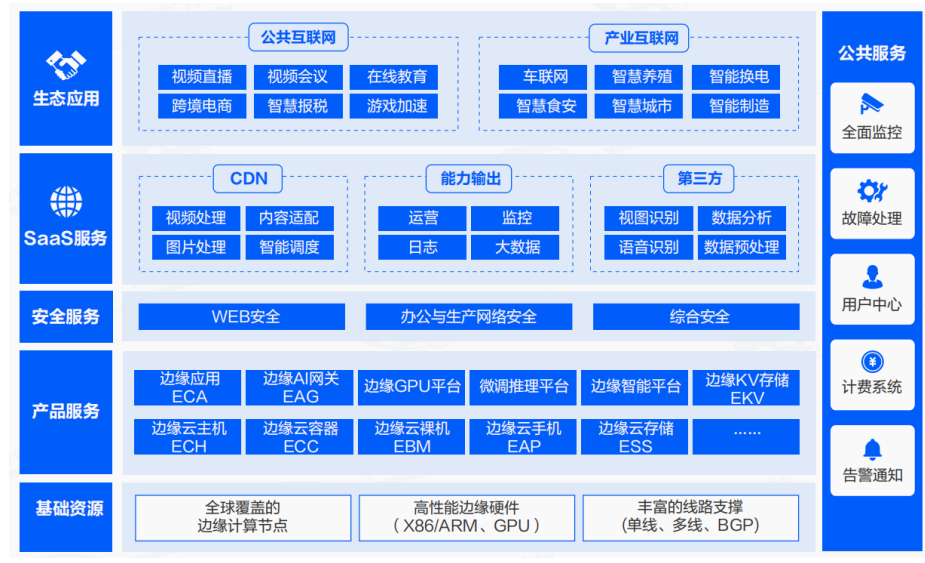


资料来源：顺网科技官方公众号，民生证券研究院

3.4 网宿科技：专注边缘计算+全球部署

专注边缘计算，积累完备技术栈。网宿科技基于深厚的 CDN 经验积累，围绕分布式、可扩展性、安全性、实时易用和可管理性等一系列核心原则，于 2018 年正式推出了高效、安全且友好的边缘计算平台 ECP。网宿科技在边缘计算平台的建设过程中投入了大量的研发资源，经历了百余次的版本迭代，沉淀了丰富的研发经验和完备的技术栈。

图14：网宿边缘计算平台架构



资料来源：网宿科技《边缘计算市场实践与洞察报告》，民生证券研究院

全球深度部署节点，边缘产品生态丰富。网宿在全球范围内建设了 2800 多个节点，遍布 70 多个国家和地区，合作覆盖国内全部运营商和海外 200+ 运营商，在线服务器超过 20 万台；平台拥有涵盖单线、多线、BGP 等多种线路类型，并支持 GPU、X86、ARM 等多种架构，丰富的算力资源储备，强大的资源服务能力，可以满足各种场景对算力和网络的需求。依托海量基础资源，网宿边缘计算平台提供丰富的产品服务，包括边缘云裸机、边缘云主机和边缘云手机等多种云服务产品形态，可以满足公共互联网和产业互联网不同客户对边缘计算的需求。网宿边缘计算的典型落地场景如在音视频实时交互场景，通过在靠近用户的边缘节点进行数据处理和传输，能够将延迟控制在毫秒级别，其中端到端国内延迟 10-50ms，国际间 100-300ms，从而确保实时互动的流畅性。

3.5 深信服：混合云架构 + 全渠道战略

“同架构混合云”解决方案，底层基于架构一致的私有云（线下）和托管云（线上），实现线上线下统一管理、统一监控运维。深信服借助“云间互联”技术安全便捷地实现线上同一 VPC 与线下大二层互通，并支持国产 X86 与 ARM 集群与线上互通；云上推出智能大脑，为线下的 IT 私有云提供远程监控和专家值守，协助用户闭环问题，提高本地数据中心的可靠性；同时建设云端服务中心，为线下私有云用户提供云上的灾备服务、数据库服务、安全防护服务等等，把复杂的 IT 运维交给云端专家。该方案广泛应用于“本地数据中心延伸”、“业务混合部署”、“混合云灾备”等场景，为企业数字化转型构筑坚实 IT 底座。

图15：深信服“同架混合云”架构



资料来源：深信服官网，民生证券研究院

坚持全渠道战略，行业客户广泛。深信服始终坚持全面渠道化战略，全组织、全流程、全行业、全业务支持伙伴发展，可匹配多元化业务场景，能充分调动各方面资源，全面覆盖市场，并建立了类型多样、结构完备的渠道生态体系，与众多优质合作伙伴共同构筑了广泛而有活力的渠道网络，有利于产品和服务在不同领域、不同区域的推广。根据深信服官网，目前深信服已服务了国内 90%以上的政府部委单位，能源行业公司产品 and 行业解决方案已服务 5000 多家电力、油气、煤炭等能源企业用户，且在安全业务方面，深信服已覆盖全国 50%以上医疗用户单位。

3.6 青云科技：混合云先行者+智算生态矩阵

混合云先行者，打通人工智能落地“最后一公里”。公司以统一的技术架构体系交付公有云、私有云、托管云等多种云模式，通过 QingCloud 专线服务实现公有云、私有云、数据中心之间的网络直连与混网组合，帮助企业构建混合云架构。青云科技 CEO 林源在 2024 年 10 月 9 日主旨演讲中阐释，青云的使命是降低新技术的使用门槛、加速应用场景落地，通过连接算力供给方与需求方，推动企业数智化转型，并与生态伙伴共同解决人工智能落地“最后一公里”的问题。

图16：青云科技混合云架构



资料来源：青云科技官网，民生证券研究院

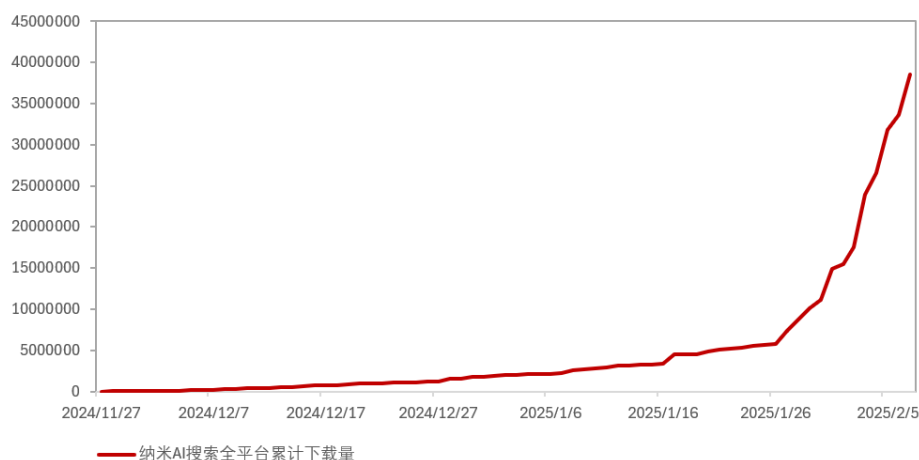
发布智算生态矩阵，智算运维先锋。青云科技已经与 200 余家生态合作伙伴紧密合作，完成了 138 项生态适配，通过算力共赢计划、算力加速计划、模型共建计划，共同构建一个开放、多元、共赢的 AI 生态体系，以资源共享与优势互补来推动智算产业的整体发展，共同加速 AI 在百行千业的落地。青云科技致力于解决智算中心建设、运维管理与运营中的挑战与痛点，公司已经落地了近 30 个区域智算中心，通过灵活的 AI 算力交付方式，青云智算中心解决方案将多个地区的算力中心统一管理、运维和运营，极大提高资源利用效率的同时，节省了大量的配置和安装时间，提高了部署的效率和准确性。

且根据青云科技官方公众号，青云科技旗下 AI 算力云服务——基石智算 CoresHub 正式上线 DeepSeek-R1 系列模型。截止 2025 年 2 月 7 日，基石智算已经部署并上线了 DeepSeek 全系列大模型供用户选择使用。

3.7 三六零：专家协作模型云协同+AI 安全护航

独创 CoE 专家协作模型技术架构，百余模型“云”上协同调动。1 月 25 日，360 纳米 AI 搜索软件-AI 搜索功能接入“DeepSeek-R1”大模型；1 月 30 日上线“DeepSeek-R1”大模型-满血高速专线版。据 360 智慧商业微信公众号，三六零独创 CoE 专家协作模型技术架构，召集了几乎行业内最强的 15 家 AI 大模型品牌，再加上 360 自研的大模型“360 智脑”品牌，共 16 家品牌的 100 多个模型作为智能底座支撑 AI 搜索的能力。根据七麦数据显示，自 2024 年 11 月 27 日纳米 AI 搜索上线以来，截止到 2025 年 2 月 7 日，纳米 AI 搜索全平台下载量从开始的 10756 极速飙升至 3850 万左右。

图17：纳米 AI 搜索全平台下载量趋势图（单位：次数）

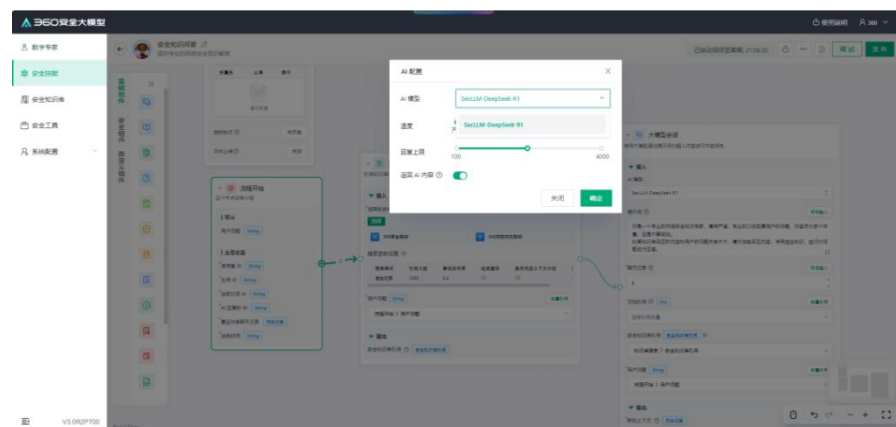


资料来源：七麦数据，民生证券研究院绘制

AI 安全为模型运行保驾护航。作为国内首家接入 DeepSeek 的数字安全企业，360 提出用 AI 重塑安全，以独创“类脑分区”专家协同（CoE）架构，发掘攻击

检测、运营处置、追踪溯源、安全知识管理、代码安全等场景，在政府、金融、央企、运营商、教育、医疗等关键基础设施行业落地使用，多次以分钟级速度帮助企业智能化拦截勒索病毒、捕获 APT 攻击，同时发挥“以模制模”的新思路，不久前审计并发现了近 40 个大模型相关安全漏洞。

图18: 360“DeepSeek 版”安全大模型



资料来源: 360 数字安全微信公众号, 民生证券研究院

3.8 金山办公：云办公行业领先者发挥新质生产力作用

全球知名的行业领先的办公软件和服务提供商。金山办公的核心产品 WPS Office 功能丰富，支持文字、表格、演示等多种办公组件，并且不断推出新的功能和服务，如金山协作、金山会议、金山日历等，形成了完整的办公协作套件矩阵。并在 2023 年发布了基于大语言模型的智能办公应用 WPS AI，锚定 AIGC（内容创作）、Copilot（智慧助理）、Insight（知识洞察）三个战略方向发展，将 AI 在国内办公软件领域率先落地的成果带给用户。截至 2023 年 12 月 31 日，公司主要产品月度活跃设备数为 5.98 亿，公司累计年度付费个人用户数达到 3,549 万，同比增长 18.43%。2024 年 8 月 5 日发布的 2024 年度《财富》世界 500 强排行榜共有 133 家中国企业上榜，其中 90% 的中国企业使用 WPS 365 实现提质增效。

发挥新质生产力作用，助力企业数字化转型。公司面向组织级客户全新升级办公新质生产力平台 WPS 365，提供内容创作、办公协作、开放生态及数字资产管理等能力。一方面，WPS 365 中的即时通讯服务、轻维表、轻审批等功能为企业提升团队协作效率、管理研发项目、低成本搭建审批流程等提供了重要支撑；另一方面，通过 WPS 365 公有云/私有化部署的云文档系统，企业数字资产被严格加密保存云端，数据安全得到有效保障，企业在数字化转型过程中的需求得到全方位满足。截至 2023 年末，公司 WPS 365 已服务 17,000 余家头部政企客户，诸如云南、浙江、山东等地党政客户，以及上交所、中信建投和科大讯飞等知名企业客户，并形成标杆效应。

图19: 金山办公生产力平台



资料来源: 金山办公微信公众号, 民生证券研究院

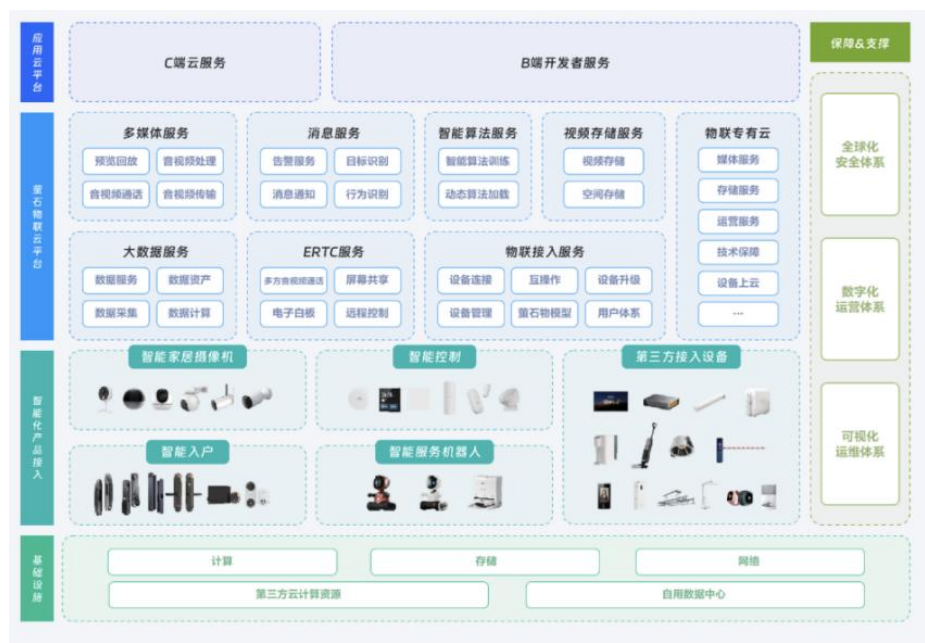
3.9 萤石网络: 以云为重, 终端+AI 的两翼齐飞

萤石以“2+5+N”生态体系, 构建其智能生活解决方案及开放式云平台服务。

萤石构建“2+5+N”智能家居生态, 以安全为核心, 以萤石云为中心, 搭载包括智能家居摄像机、智能入户、智能控制、智能服务机器人在内的四大自研硬件, 开放接入环境控制、智能影音等子系统生态, 实现家居及类家居场景的全屋智能化, 同时利用互联互通的萤石云开放平台, 与合作伙伴分享智能视频的云平台服务能力, 共同打造物联网云生态。自2020年公司IoT物联云平台发布当年即平台接入设备数突破1亿, 平台月活用户突破2600万。

萤石以云为核心, 云端闭环, 共谋协同增长。公司通过租赁数据中心和采购第三方云计算资源的方式, 打造了计算、存储、网络等云基础设施保障; 基于底层IaaS资源, 公司自主研发建设PaaS层物联云平台, 包括音视频多媒体、消息通知处理、智能算法调度、视频存储备份、ERTC、大数据、物联接入等多种云平台PaaS服务。目前萤石云已成为全球领先的视频/视觉公有云平台, 作为全球化物联云服务平台, 通过构建多数据中心+就近服务点的方式服务于全球客户。

图20：萤石网络业务架构图



资料来源：萤石网络 2022 年年报，民生证券研究院

端侧 AI 算力需求同样契合萤石产品矩阵属性和发展战略。萤石网络目前已构建摄像头、智能门锁、清洁机器人等多个热门终端，并以立身之本的萤石云为基石，通过 AI 全面赋能，实现多终端并发叠加 AI 赋能。

3.10 软通动力：天璇 AI 平台获 DeepSeek 优化能力跃迁

DeepSeek-R1 模型优化天璇，跃升获得“知识炼金”能力。作为中国数字技术产品和服务创新领导企业，软通动力 2 月 4 日宣布积极拥抱 DeepSeek，率先进行产品的创新融合，通过把 DeepSeek-R1 接入天璇 MaaS 平台，以全栈 AI 技术服务加速企业智能化转型。天璇 MaaS 平台是向客户提供一站式的企业大模型技术底座，支持客户开发并管理行业大模型和应用场景大模型，已经在银行、保险、零售、农业、医疗、通用管理等多个垂直行业领域赋能客户智能化创新。平台始终秉持汇聚前沿 AI 模型的理念，全力加速创新进程，为客户提供卓越服务。同时，软通咨询则从战略高度出发，助力企业落地 AI 应用。通过深入的大模型分析、科学的数据治理、精准的模型生态与技术路线选择，不仅能够创新商业范式、提升咨询效率，还能切实提升企业的业务价值与客户满意度，引领行业迈向智能新高度。

软通动力通过整合 DeepSeek-R1 模型并进一步优化天璇 AI Foundation 平台，凭借其跨行业的深厚积累和强大的硬件生态优势，致力于提供安全、可靠的高性能 AI 解决方案，支撑企业在生产推理和训练方面的高效率需求，实现行业模型

的快速落地,以低成本、高性能的方式推动个性化 AI 应用的发展。借助 DeepSeek-R1 模型在语言理解和推理方面的卓越能力,软通动力的天璇知识库实现了从被动检索向主动构建动态知识架构的跃升,获得了“知识炼金”的能力。

图21: DeepSeek-R1 接入天璇 MaaS 平台



资料来源:软通动力个官方公众号,民生证券研究院

智能化是软通动力的重要发展战略,基于天璇智脑中枢构建了软硬全链条 AI 产品及服务。以咨询为牵引,融合如 DeepSeek-R1 等卓越模型,持续进行技术与产品创新,助推企业大模型加速落地,并基于自身强大的技术生态搭建人工智能软硬件协同(包含训推一体机、AIPC、人形机器人、复合机器人等)产业链条,为客户提供全栈式高质量 AI 解决方案。

3.11 科大讯飞: 讯飞星火深耕 AI 教育领域

基于全国算力训练的国内领先讯飞星火大模型。2023 年 10 月 24 日,科大讯飞与华为联合发布了国内首个全国算力平台“飞星一号”。2024 年 10 月 24 日,基于全国首个国产万卡算力集群训练的全民开放大模型讯飞星火 4.0 Turbo 正式发布。

根据真实数据背靠背的测试,七项核心能力在中文领域全面超过 GPT-4 Turbo,代码能力和数学能力超越 GPT-4o 且星火 4.0 Turbo 在行业能力上也有明显的提升,例如,金融领域知识问答绝对提升 14%,油气领域绝对提升了 16%;在艾伦人工智能研究所、OpenAI 等权威机构发布的 14 项主流测试集中,讯飞星火 4.0 Turbo 实现对美国三大主流模型(GPT-4o、Claude 3.5 Sonnet、Gemini 1.5pro)的 9 项超越,效率相对提升 50%。

图22：讯飞星火 4.0 Turbo 行业能力全面提升



资料来源：科大讯飞官方公众号，民生证券研究院

深耕 AI 教育领域，智慧办公 SaaS 平台业务广覆盖。科大讯飞的智慧教育业务实现了教育的全场景覆盖，包括教学、学习、考试、管理等，目前已覆盖全国 32 个省级行政区，超过 5 万所学校，累计服务师生超过 1.3 亿，并在日本、新加坡等海外市场应用。2024 年发布的 AI 科学教育解决方案发布，已覆盖 5 万余名学生；星火智能批阅机在全国 260 多所学校开展试点应用；星火教师助手累计覆盖全国 2000 余所学校 10 万余名教师，让教学设计效率提升 56%+、资源检索便捷度提升 56%+、课件制作效率提升 64%+。2024 年“讯飞星火”App 安卓下载量达 2 亿，讯飞智慧办公 SaaS 平台累计覆盖用户超 8000 万，生态用户 2 亿+。

4 风险提示

1) 技术发展不确定性。云计算技术快速迭代，云厂商需要持续投入大量资金进行技术研发和设备更新。如果技术发展不及预期，可能导致云厂商在市场竞争中处于劣势。过度依赖第三方开源模型可能削弱云厂商的技术自主性。如果开源模型的技术发展停滞或出现兼容性问题，云厂商可能面临技术瓶颈。

2) 行业竞争加剧风险。随着基于 DeepSeek 系列模型提供的服务同质化严重，云厂商可能会为吸引用户易引发“价格战”，且从长远来看，如果行业普遍采用 DeepSeek 的低成本模式，云服务可能从“高毛利闭源产品”转向“低毛利开源服务”，削弱云厂商的盈利能力。

插图目录

图 1: MLA 及 DeepSeekMOE 基础架构.....	3
图 2: DeepSeek-R1 系列模型性能对比.....	4
图 3: 蒸馏的技术原理.....	4
图 4: DeepSeek 蒸馏小模型的性能测评.....	5
图 5: 成都华微 R1 蒸馏测试代码.....	6
图 6: 超级产品增长 1 亿用户所用的时间.....	8
图 7: 中国云计算市场规模及增速(单位: 亿元).....	8
图 8: 金山云服务器获取 DeepSeek 镜像.....	9
图 9: 算力平权下的云商业模式闭环逻辑.....	13
图 10: 海康威视国际业务布局一览.....	13
图 11: 2023 年中国央企云“IaaS+PaaS”市场竞争格局.....	15
图 12: 优刻得算力共享平台.....	16
图 13: “顺网算力”业务架构.....	17
图 14: 网宿边缘计算平台架构.....	18
图 15: 深信服“同架混合云”架构.....	19
图 16: 青云科技混合云架构.....	19
图 17: 纳米 AI 搜索全平台下载量趋势图 (单位: 次数).....	20
图 18: 360“DeepSeek 版”安全大模型.....	21
图 19: 金山办公生产力平台.....	22
图 20: 萤石网络业务架构图.....	23
图 21: DeepSeek-R1 接入天璇 MaaS 平台.....	24
图 22: 讯飞星火 4.0 Turbo 行业能力全面提升.....	25

表格目录

表 1: DeepSeek 系列模型硬件需求.....	6
表 2: 国内公有云厂商接入 DeepSeek 模型.....	11
表 3: 云服务模式及相关代表公司.....	12

分析师承诺

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并登记为注册分析师，基于认真审慎的工作态度、专业严谨的研究方法与分析逻辑得出研究结论，独立、客观地出具本报告，并对本报告的内容和观点负责。本报告清晰地反映了研究人员的研究观点，结论不受任何第三方的授意、影响，研究人员不曾因、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。

评级说明

投资建议评级标准	评级	说明
以报告发布日后的 12 个月内公司股价（或行业指数）相对同期基准指数的涨跌幅为基准。其中：A 股以沪深 300 指数为基准；新三板以三板成指或三板做市指数为基准；港股以恒生指数为基准；美股以纳斯达克综合指数或标普 500 指数为基准。	推荐	相对基准指数涨幅 15%以上
	谨慎推荐	相对基准指数涨幅 5% ~ 15%之间
	中性	相对基准指数涨幅-5% ~ 5%之间
	回避	相对基准指数跌幅 5%以上
行业评级	推荐	相对基准指数涨幅 5%以上
	中性	相对基准指数涨幅-5% ~ 5%之间
	回避	相对基准指数跌幅 5%以上

免责声明

民生证券股份有限公司（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。

本报告仅供本公司境内客户使用。本公司不会因接收人收到本报告而视其为客户。本报告仅为参考之用，并不构成对客户的投资建议，不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告所包含的观点及建议并未考虑获取本报告的机构及个人的具体投资目的、财务状况、特殊状况、目标或需要，客户应当充分考虑自身特定状况，进行独立评估，并应同时考量自身的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见，不应单纯依靠本报告所载的内容而取代自身的独立判断。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而导致的任何可能的损失负任何责任。

本报告是基于已公开信息撰写，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、意见及预测仅反映本公司于发布本报告当日的判断，且预测方法及结果存在一定程度局限性。在不同时期，本公司可发出与本报告所刊载的意见、预测不一致的报告，但本公司没有义务和责任及时更新本报告所涉及的内容并通知客户。

在法律允许的情况下，本公司及其附属机构可能持有报告中提及的公司所发行证券的头寸并进行交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问、咨询服务等相关服务，本公司的员工可能担任本报告所提及的公司的董事。客户应充分考虑可能存在的利益冲突，勿将本报告作为投资决策的唯一参考依据。

若本公司以外的金融机构发送本报告，则由该金融机构独自为此发送行为负责。该机构的客户应联系该机构以交易本报告提及的证券或要求获悉更详细的信息。本报告不构成本公司向发送本报告金融机构之客户提供的投资建议。本公司不会因任何机构或个人从其他机构获得本报告而将其视为本公司客户。

本报告的版权仅归本公司所有，未经书面许可，任何机构或个人不得以任何形式、任何目的进行翻版、转载、发表、篡改或引用。所有在本报告中使用的商标、服务标识及标记，除非另有说明，均为本公司的商标、服务标识及标记。本公司版权所有并保留一切权利。

民生证券研究院：

上海：上海市浦东新区浦明路 8 号财富金融广场 1 幢 5F； 200120

北京：北京市东城区建国门内大街 28 号民生金融中心 A 座 18 层； 100005

深圳：深圳市福田区中心四路 1 号嘉里建设广场 1 座 10 层 01 室； 518048