证券研究报告/行业研究

报告日期：2024 年 12 月 31 日

豆包大模型迎来重磅更新，AI 应用繁荣推动算力基础设施建设

——人工智能行业专题研究

投资要点

➢ 豆包大模型重磅更新，有望带动AI应用市场繁荣

2024年12月火山引擎冬季FORCE原动力大会推出豆包视觉理解大模型和3D生产大模型，并对通用模型Pro、音乐生产模型和文生图模型性能升级。根据智源研究院发布的FlagEval“百模”评测结果，豆包通用模型Pro在大语言模型总榜中主观评测得分最高，豆包视觉理解模型和文生图模型在多模态模型主观测评中得分第二。豆包视觉理解大模型和3D生产大模型具备的内容识别、视觉描述和3D内容生成能力进一步增强AI应用实用性，而使用成本相比行业价格可降低85%，有望推动AI应用市场的商业繁荣。

➢ 大模型的大规模商业化应用已成熟，拉动算力基础设施建设

人工智能行业已跨过AI大模型大规模成熟商业化应用的节点，国内厂商加大对AI Agent等新一代人工智能应用的投入。AI大模型性能提升所需的千亿级参数训练及应用端繁荣对算力规模的需求，都将推动算力基础设施的建设。根据IDC数据，2024年全球人工智能资本开支有望达2350亿美元，并预计2028年增长至6320亿美元，复合增速达29%。此外生成式人工智能资本开支2024-2028年GAGR有望达59%，显著高于其他人工智能技术的22%。

➢ 算力基础设施建设趋势下，核心供应链环节将充分受益

人工智能行业带动算力基础设施建设趋势下，服务器、液冷设备、芯片及光模块等是供应链的核心环节。1）服务器：服务器是算力载体，AI服务器比普通服务器对算力及存储性能要求更高，2024年全球普通AI服务器和高端AI服务器出货量分别为72.5和54.3万台，分别同比增长54.2%和172.0%。2）液冷设备：液冷服务器具有低能耗、高散热优点，符合高算力数据中心需求；3）芯片：芯片是算力大脑，卡脖子风险最高，国产芯片亟待突破。4）光模块：800G和1.6T高端光模块用量有望大幅提升，国产公司在全球市场具有领先地位。

➢ 投资建议

豆包大模型产品力大幅提升，并大幅降低人工智能大模型使用成本，有望推动AI应用的商业繁荣。建议关注服务器、液冷设备、芯片和光模块等领域的投资机会：1）服务器：浪潮信息、中科曙光；2）液冷设备：英维克；3）芯片：海光信息；4）光模块：中际旭创、天孚通信、光迅科技。

➢ 风险提示

AI应用渗透不及预期；算力开支不及预期；宏观经济不及预期；竞争加剧。

投资评级： 看好

分析师：吴起涤

执业登记编号：A0190523020001

wuqidi@yd.com.cn

研究助理：程治

执业登记编号：A0190123070008

chengzhi@yd.com.cn

人工智能指数与沪深 300 指数走势对比

资料来源：Wind，源达信息证券研究所

1.《人工智能专题研究系列一：大模型推动各行业AI应用渗透》2023.08.02
2.《人工智能专题研究系列二：AI大模型开展算力竞赛，推动AI基础设施建设》2023.08.03
3.《人工智能专题研究系列三：Gemini1.0有望拉动新一轮AI产业革新，算力产业链受益确定性强》2023.12.12
4.《人工智能专题研究系列四：OpenAI发布Sora文生视频模型，AI行业持续高速发展》2024.02.19
5.《人工智能专题研究系列五：Kimi智能助手热度高涨，国产大模型加速发展》2024.03.26
6.《数据中心液冷技术专题研究：算力扩建浪潮下服务器高密度、高耗能特征显著，催动液冷技术市场快速扩容》2024.08.15

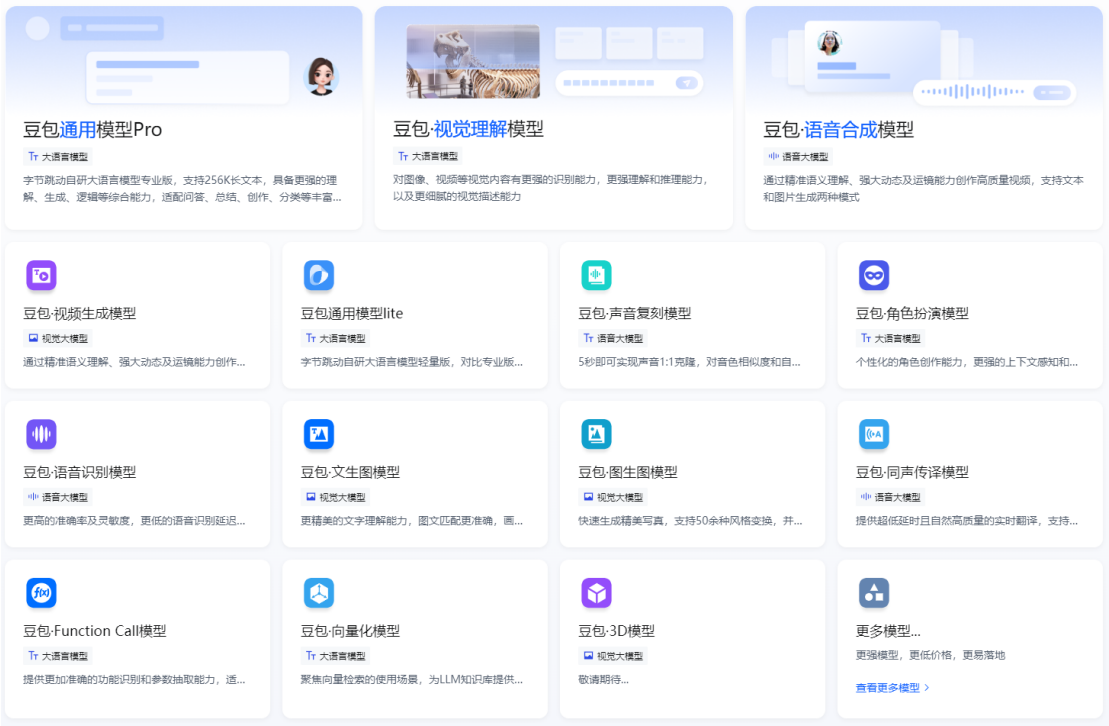
目录

图表目录

一、豆包大模型产品力大幅增强，推动 AI 应用商业繁荣

2024 年 12 月火山引擎冬季 FORCE 原动力大会推出豆包视觉理解大模型和 3D 生产大模型，并将通用模型 Pro、音乐生产模型和文生图模型升级，进一步丰富产品矩阵。2024 年 5 月火山引擎春季 FORCE 原动力大会首次发布豆包大模型系列产品以来，仅 7 个月就再度升级，并在多项大模型能力测评榜单中居于前列。本次大会新推出的豆包视觉理解大模型和 3D 生产大模型拓展了模型的内容识别、视觉描述和 3D 内容生成能力，并显著降低使用成本，有望推动人工智能应用端的商业繁荣。

图 1：豆包大模型产品矩阵丰富



资料来源：火山引擎，源达信息证券研究所

豆包视觉理解模型具备更强内容识别能力和理解推理能力。豆包视觉理解模型具备强大的图片理解与推理能力及精准的指令理解能力。模型在图像文本信息抽取、基于图像的推理任务上有展现出了强大的性能，能够应用于更复杂、更广泛的视觉问答任务。比如模型可描述图片内容并根据图片内容进行提问。此外，该模型可完成深度的图片理解与推理，在表格图像、数学问题、代码图像等复杂推理场景下完成任务。

图 2：豆包视觉理解模型具备更强内容识别能力

资料来源：火山引擎，源达信息证券研究所

图 3：豆包视觉理解模型具备更强理解和推理能力

资料来源：火山引擎，源达信息证券研究所

火山引擎冬季 FORCE 原动力大会首次发布豆包 3D 生成模型。该模型可支持文生 3D、图生 3D 及多模态生成等诸多功能，模型与火山引擎数字孪生平台 veOmniverse 结合使用，可高效完成智能训练、数据合成和数字资产制作，成为一套支持 AIGC 创作的物理世界仿真模拟器。

图 4：火山引擎首次发布豆包 3D 生成模型

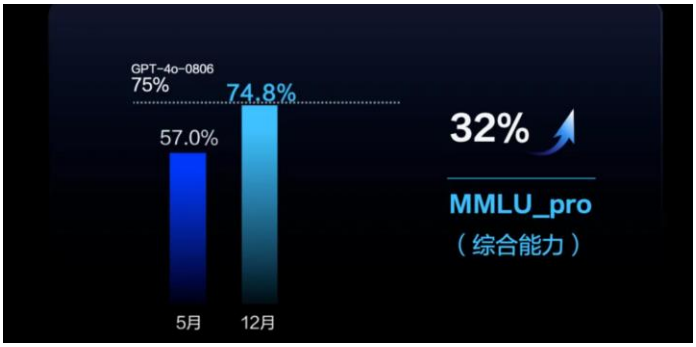
资料来源：火山引擎，源达信息证券研究所

图 5：豆包 3D 生成模型可根据文本生成 3D 场景

资料来源：火山引擎，源达信息证券研究所

火山引擎对豆包通用模型 Pro 进行升级，模型性能大幅提升。豆包通用模型 Pro 相比 2024 年 5 月发布版本，在综合能力上提升 32%，与 GPT-4o 持平，而使用成本仅是其八分之一。模型在指令遵循、代码、专业知识、数学层面对齐 GPT-4o 水平，其中指令遵循能力提升 9%，代码能力提升 58%，GPQA 专业知识方面能力提升 54%，数学能力提升 43%，推理能力提升 13%。

图 6：豆包通用模型 Pro 综合能力大幅提升



资料来源：火山引擎，源达信息证券研究所

图 7：通用模型 Pro 在指令遵循、代码、数学等指标对标 GPT-4o

资料来源：火山引擎，源达信息证券研究所

火山引擎对豆包文生图模型和音乐模型能力升级。1）豆包文生图模型：模型在通用性、可控性、高质量方面实现突破，并新增一键海报和一键 P 图功能，可根据用户简单指令对图片进行精准编辑，并加强了对文字细节的指令遵循能力。2）豆包音乐模型：可根据用户简单描述或上传图片，生成时长 3 分钟的包含旋律、歌词和演唱等元素在内的音乐作品，包括前奏、主歌、副歌、间奏、过渡段等复杂结构，并支持局部修改功能，在针对部分歌词修改后仍能在原有旋律的节奏框架内适配。

图 8：豆包文生图模型能力升级

资料来源：火山引擎，源达信息证券研究所

图 9：豆包音乐模型能力升级

资料来源：火山引擎，源达信息证券研究所

豆包大模型能力显著提升，在多项能力测评中排名前列。根据 2024 年 12 月 19 日智源研究院发布的 FlagEval“百模”评测结果，在闭源大模型评测能力总榜中，豆包通用模型 Pro 在大语言模型总榜中主观评测得分最高；多模态模型评测总榜中，豆包视觉理解模型和文生图模型的主观测评得分均排名第二。

图 10：豆包通用模型 Pro 在大模型测评总榜中排名第一

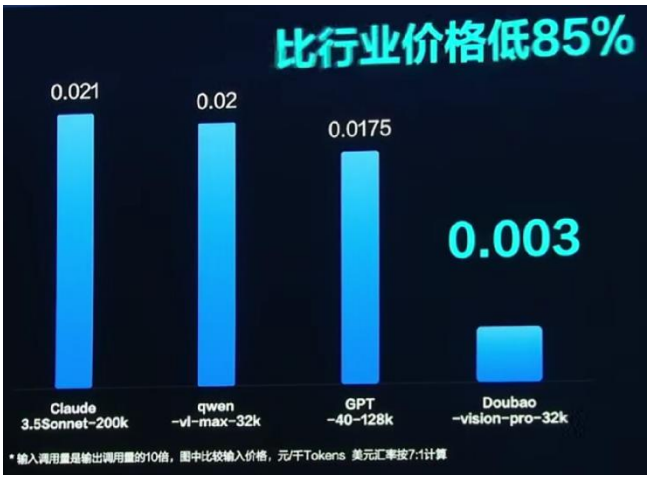
资料来源：智源研究院，源达信息证券研究所

图 11：豆包视觉理解模型在视觉语言模型测评榜单中排名第二

资料来源：智源研究院，源达信息证券研究所

豆包视觉理解模型使用成本大幅低于行业平均水平，有望推动 AI 应用商业化成熟。根据火山引擎冬季 FORCE 原动力大会数据，豆包视觉理解模型的使用成本是 0.003 元/千 Tokens，大幅低于 GPT-4o 的 0.0175 元/千 Tokens，比行业价格低 85%，大模型使用成本降低有望推动 AI 应用商业化。根据 AI 产品榜数据，豆包 APP 在 2024 年 11 月全球 AI 产品榜中排名第二，在终端应用的渗透率进一步提升。

图 12：豆包视觉理解模型使用成本大幅低于行业平均水平



资料来源：火山引擎，源达信息证券研究所

图 13：豆包 APP 在 11 月全球 AI 产品榜中排名第二

1	AI产品榜 aicpb.com · 全球总榜					
2	全球排名		产品名 AI产品榜	应用(APP)简短描述 aicpb.com	11月上榜应用 APP MAU	11月上榜应用 MAU变化
3	1		ChatGPT	The official app by OpenAI	287.25M	11.27%
4	2		豆包	AI 智能助手 | 抖音	59.98M	16.92%
5	3		Nova	聊天AI与AI写作机器人	49.63M	5.67%
6	4		ChatOn	Powered by ChatGPT & GPT-4o	28.84M	6.66%
7	5		Remini	人工智能修图	27.96M	-2.16%
8	6		Character AI	Chat Ask Create	26.88M	5.74%

资料来源：智源研究院，源达信息证券研究所

二、人工智能产业加快增长，应用及算力是两大支柱

AI 大模型对算力需求大，推动 AI 基础设施建设。AIGC 行业进入高速发展期，AI 大模型性能持续提升的背后是千亿级以上的参数训练，带来对算力的高额需求，有望推动新一轮 AI 基础设施建设。根据 OpenAI 官网，AI 模型训练计算量自 2012 年起每 3.4 个月就增长一倍。以 GPT-3 模型为例，根据 lambdalabs 数据，该模型参数规模达 1750 亿，完整训练运算量达 3640PFlop/s-days（以 3640PFlop/s 速度进行运算，需要 3640 天）。模型完成单次训练约需要 355 个 CPU 年并耗费 460 万美元（假设采用 Nvidia Tesla V100 芯片）。

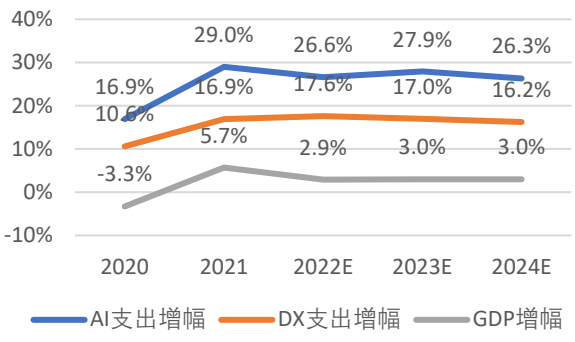
表 1：人工智能大模型的参数规模呈指数级增长趋势

Models	Release time	Developers	Parameter size/10^{-8}	Sample size/10^{-9}
GPT-1	2018	OpenAI	1.17	10
BERT	2018	Google	3.40	34
GPT-2	2019	OpenAI	15.00	100
Fairseq	2020	Meta	130.00	—
GPT-3	2020	OpenAI	1750.00	4990
GLaM	2021	Google	1200.00	16000
LaMDA	2022	Google	1370.00	15600
GPT-4	2023	OpenAI	—	—
Ernie Bot	2023	Baidu	—	—
SparkDesk	2023	iFLYTEK	1700.00	—
PanguLM	2023	HUAWEI		>30000

资料来源：《大语言模型研究现状及趋势》，源达信息证券研究所

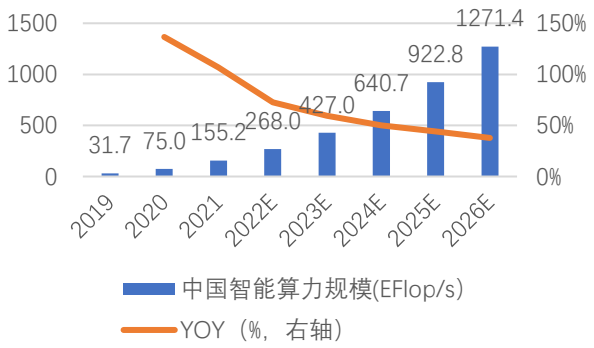
高算力需求迫切，推动 AI 基础设施建设。高训练算力需要与相应基础设施匹配，根据《2022-2023 中国人工智能计算力发展评估报告》预计，2024 中国智能算力规模将达 641EFlop/s，同比增长 50%，并预计 2025 年中国智能算力将达 923Eflop/s，同比增长 44%。。

图 14：预计 2022-2024 年全球 AI 支出年增速高于 20%



资料来源：IDC，世界银行，源达信息证券研究所

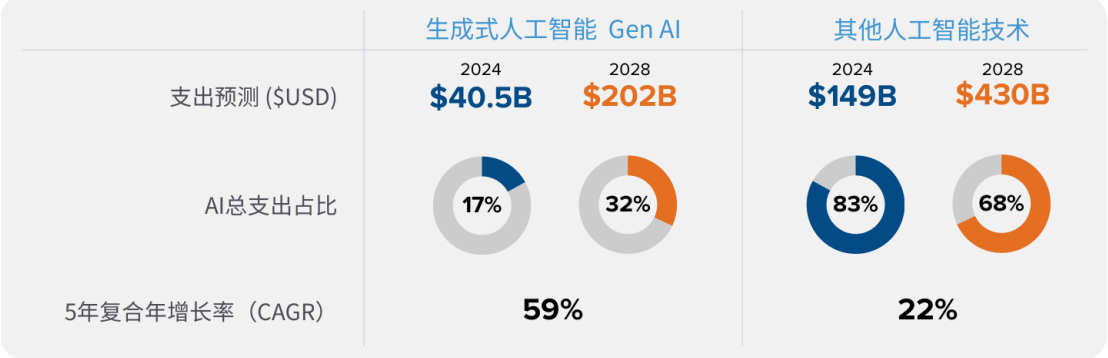
图 15：预计 2024 年中国智能算力规模同比增长 50%



资料来源：IDC，源达信息证券研究所

IDC 预计 2024 年全球人工智能资本开支达 2350 亿美元，并预计 2028 年增长至 6320 亿美元，复合增速达 29%。此外生成式人工智能资本开支 2024-2028 年 GAGR 有望达 59%，显著高于其他人工智能技术的 22%。

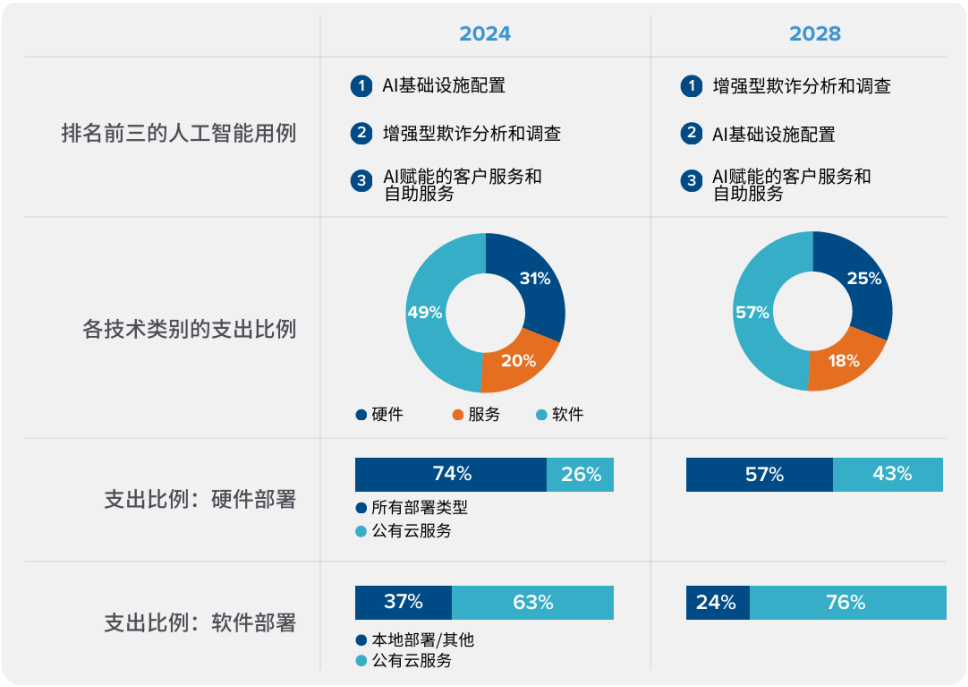
图 16：IDC 预计 2024-2028 年全球人工智能资本开支复合增速 GAGR 达 29%



资料来源：IDC，源达信息证券研究所

根据 IDC 数据，人工智能支出排名前三的行业是软件和信息服务、银行及零售业，预计在 2024 年的 AI 投资达 896 亿美元，占全球市场的 38%。银行业将显著增加对欺诈分析和调查等 AI 服务的需求。而软件开支未来在人工智能支出中占比最高，预计将显著带动 IAAS、SAAS、PAAS 等云端服务市场的增长。

图 17：IDC 预计 2028 年软件资本开支将占人工智能支出的 57%



资料来源：IDC，源达信息证券研究所

大模型加速发展趋势下，国内厂商加大对 AI Agent 等新一代人工智能应用的投入。AI Agent 是一种以 AI 大模型驱动的人工智能工具，可根据具体场景实现高度个性化和智能化的智能服务，有望将大模型的潜力最大化，推动 AI 技术应用化，加速人工智能产业商业化。

表 2：国内厂商加大对 AI Agent 等大模型驱动下的人工智能应用的投入

公司名称	大模型产品
阿里云	百炼大模型服务平台
AWS	Amazon bedrock 以及 partyrock.aws 等工具
百度智能云	TiAppBuilder、AgentBuilder
京东云	AI Agent 开发管理平台
蚂蚁集团/蚂蚁数科	蚂蚁 AI Studio+Max
昆仑万维	SkyAgents
商汤科技	MaaS 平台-应用智能体
深信服科技	AI 算力平台
神州数码	神州问学-AI 应用及 Agent 管理
腾讯云	腾讯元器
月之暗面	Kimi Plus
中国电信(天翼 AI)	智能体开发运营平台
字节跳动	火山引擎 AI Agent 开发管理平台、豆包 APP
360	360 智脑、360 智汇云

资料来源：IDC，源达信息证券研究所

三、算力产业链：服务器是算力基础设施

1.大模型打开算力需求，服务器建设规模快速增长

大模型发展打开算力需求，AI 算力服务器需求有望增长。自 OpenAI 发布 ChatGPT 后，AI 大模型有望成为助力千行万业智能化转型的底层支撑。AI 大模型的训练和运行过程对对算力需求极大，预计将推动一轮算力中心的建设。以 Nvidia A100 服务器为例（由 8 个 A100 GPU 构成），单台服务器算力约为 5Pflop/s，则训练一个具有 1750 亿个模型参数的大模型需要约 2917 台 A100 服务器。

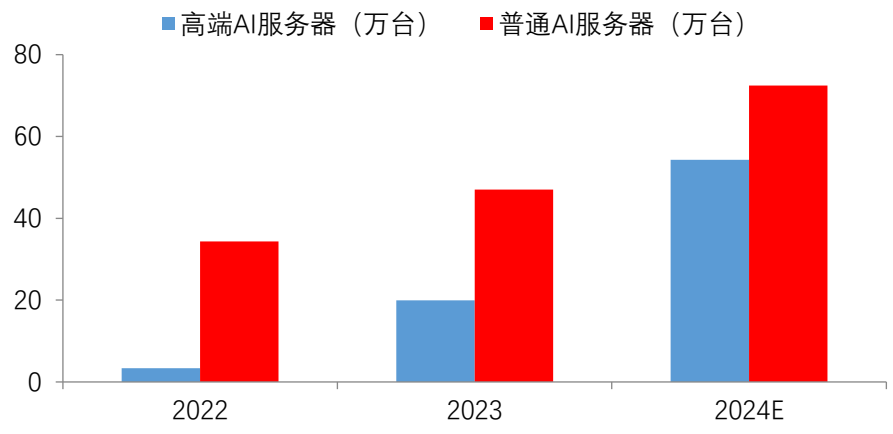
表 3：具有 1750 亿个模型参数的大模型训练一天需要约 2917 台 Nvidia A100 服务器

模型参数（亿个）	350	700	1050	1400	1750
所需算力（E+8PFlop/s）	0.63	1.26	1.89	2.52	3.15
有效算力比率（%）	25%	25%	25%	25%	25%
实际算力需求（E+8PFlop/s）	2.52	5.04	7.56	10.08	12.6
服务器算力（PFlop/s）	5	5	5	5	5
每日工作时间（s）	86400	86400	86400	86400	86400
服务器需求数（台）	583	1167	1750	2333	2917

资料来源：Nvidia 官网，OpenAI，源达信息证券研究所

人工智能行业高速发展，算力巨额缺口推动 AI 服务器出货量高速增长。2023 年全球普通 AI 服务器/高端 AI 服务器出货量分别为 47.0 和 27.0 万台，较 2022 年分别同比增长 36.6%和 490.5%，并预计 2024 年全球普通 AI 服务器和高端 AI 服务器出货量分别为 72.5 和 54.3 万台，分别同比增长 54.2%和 172.0%。

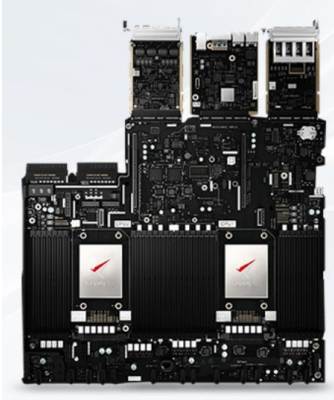
图 18：AI 服务器出货量高速增长



资料来源：华勤技术投资者关系公众号，源达信息证券研究所

华为加大算力基础设施研发力度。目前华为算力基础设施布局中：鲲鹏系列以通用算力为主，昇腾系列以智能算力为主，均采用国产芯片打造。华为凭借自身强大的研发能力，已实现从算力、存力、互联技术和计算架构等方面为世界提供第二选择，打造算力坚实底座。从产业链布局看，目前华为主要负责服务器或其中核心器件的研发和生产，并由下游服务器厂商代理销售，主要的华为系服务器厂商有高新发展（对华鲲振宇持股 70%）、四川长虹、神州数码、拓维信息和烽火通信等。此外 2023 年 3 月中兴通讯宣布自身服务器将为百度“文心一言”提供算力支撑。

图 19：搭载鲲鹏 920 处理器的鲲鹏服务器主板



资料来源：华为官网，源达信息证券研究所

图 20：华为推出昇腾系列 AI 算力基础设施

资料来源：华为官网，源达信息证券研究所

2.液冷技术低能耗高散热，受益算力扩建浪潮

液冷技术相较于传统风冷技术，具有低能耗、高散热、低噪声和低 TCO 等优点，符合数据中心高能耗、高密度的发展趋势：

1）高效散热：液体的冷却能力是空气的 1000-3000 倍，使得液冷技术更适用于高能耗、高功率的服务器应用场景。

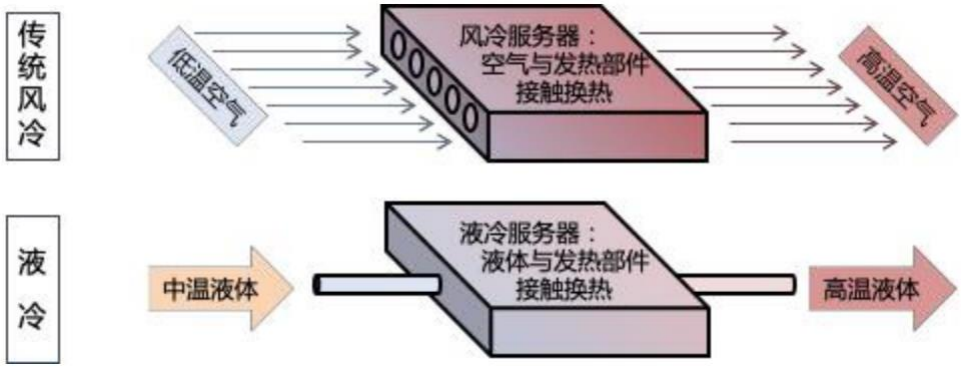
2）节能降耗：液冷系统可实现更高能效比，降低数据中心能耗。液冷技术（尤其是浸没式液冷）可将数据中心的 PUE 值降至 1.2 以下，相较于传统风冷技术，可以节省电量 30~50%。

3）提高设备可靠性：液冷技术可以减少因高温引起的设备故障，延长服务器的使用寿命，并避免因风扇引起振动和噪音。

4）节省空间：液冷技术允许更紧凑的服务器布局，无需像风冷那样需要较大的空气流通空间，从而节省了数据中心的占地面积。

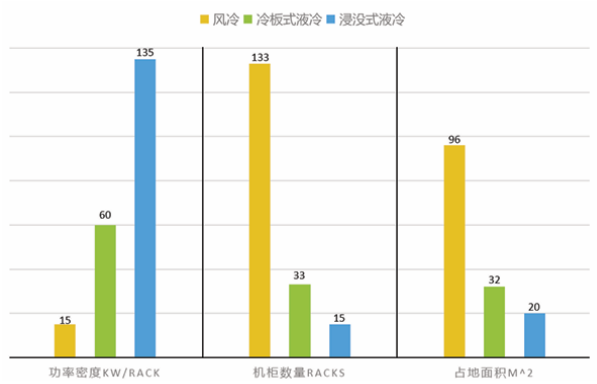
5）提高功率密度：液冷技术可以支持更高的机架功率密度，满足高性能计算和 AI 应用的需求。浸没式液冷方案可以将单机架功率提升到 100kW 甚至 200kW 以上。

图 21：传统风冷技术与液冷技术原理对比



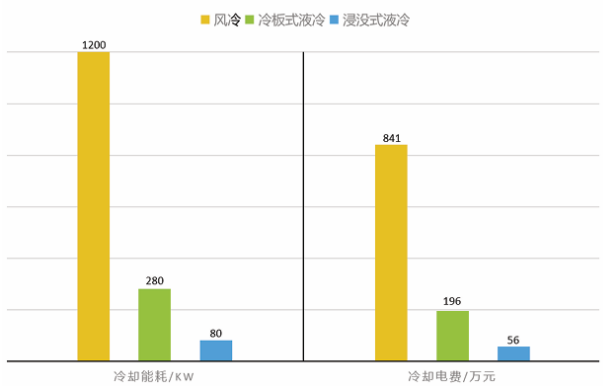
资料来源：曙光数创招股说明书，源达信息证券研究所

图 22：液冷技术散热能力显著优于风冷技术



资料来源：《中兴液冷技术白皮书》，源达信息证券研究所

图 23：液冷技术节能水平显著优于风冷技术



资料来源：《中兴液冷技术白皮书》，源达信息证券研究所

冷板式和浸没式等主流液冷技术在散热性、集成度、能效等冷却效果指标上显著优于传统风冷技术。

表 4：主流液冷技术与传统风冷技术冷却效果指标对比

	传统风冷	冷板液冷	浸没单相液冷	浸没相变液冷
散热性能	0	+	+	++
集成度	0	+	+	++
可维护性	0	+	+	+
可靠性	0	+	+	+
性能	0	+	+	++
能效	0	+	+	++
废热回收	0	+	+	++
噪音	0	++	+	++

资料来源：曙光数创招股说明书，源达信息证券研究所

人工智能变革和数字经济转型趋势下，数据中心往高能耗、高密度方向发展，液冷技术应用渐广。传统的风冷方式只能满足 2.7kW/机柜的数据中心散热需求，无法适应中高密度数据中心需求。液冷技术利用高比热容的特点和对流传热的能力，可以满足 2.7-30kW/机柜的数据中心散热需求，解决超高热流密度的散热问题，未来液冷技术必将在数据中心领域得到愈加广泛的应用。

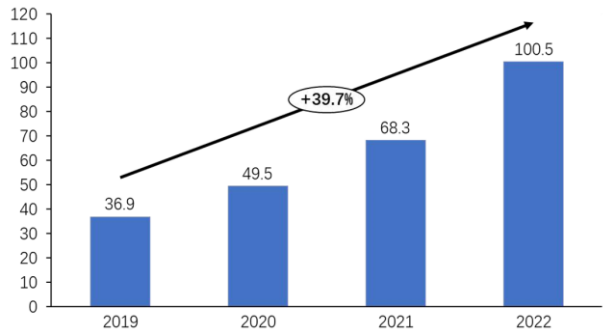
表 5：不同密度数据中心适用的冷却技术

每平方功率	数据中心密度	制冷方式
1.2Kw/机柜以下	超低密度数据中心	风冷
1.2-2.7kW/机柜	低密度数据中心	风冷
2.7-7.5kW/机柜	中、低密度数据中心	风冷/液冷
7.5-18kW/机柜	中、高密度数据中心	冷板式液冷
18-30kW/机柜	高密度数据中心	冷板式液冷/浸没式液冷

资料来源：曙光数创招股说明书，源达信息证券研究所

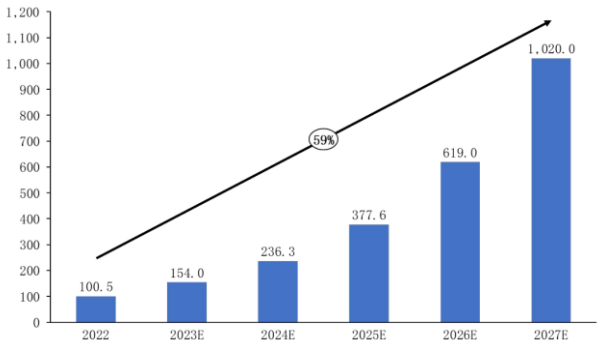
根据 2023《中国液冷数据中心市场深度研究报告》，预计 2025 年中国液冷数据中心市场规模有望达 377.6 亿元，同比增长 56%。基于市场需求发展及产业生态建设进程，未来五年中国液冷数据中心市场将以 59%的复合增长率持续发展。预计到 2027 年，AI 大模型商用落地，液冷生态趋于成熟，市场规模将出现较大幅度增长，有望达到 1020 亿元。

图 24：2019-2022 年中国液冷数据中心市场规模



资料来源：《中国液冷数据中心市场深度研究报告》，源达信息证券研究所

图 25：2022-2027 年中国液冷数据中心市场规模预测

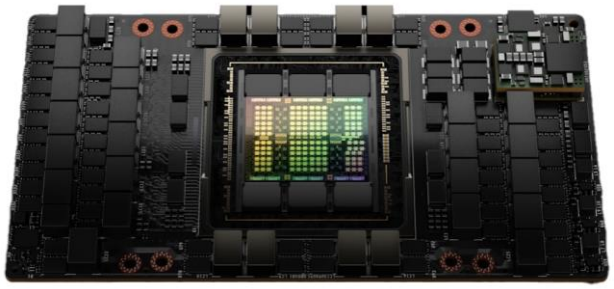


资料来源：《中国液冷数据中心市场深度研究报告》，源达信息证券研究所

四、算力产业链：芯片是智能核心，国产化短板明显

Nvidia H200 是目前最先进的人工智能芯片。2023 年 11 月 13 日 Nvidia 推出新款人工智能芯片 GPU H200，与公司上一代产品 H100 相比在存储性能上得到大幅提升，而在算力层面性能指标未有显著改变。

图 26：Nvidia GPU H200 芯片示意图



资料来源：Nvidia，源达信息证券研究所

图 27：H200 较 H100 相比在存储性能上有大幅提升

	H200 SXM	H100 SXM
Memory Clock	6.5Gbps HBM3E	5.24Gbps HBM3
Memory Bus width	6144-bit	5129-bit
Memory Bandwidth	4.8TB/sec	3.35TB/sec

资料来源：Nvidia，源达信息证券研究所

国产 AI 芯片短板明显，下一代产品推进顺利。我们通过对国内寒武纪、华为昇腾和沐曦等国产公司旗下的 AI 旗舰芯片与 Nvidia H100 SXM 的性能指标对比，可以看到国产 AI 芯片与 Nvidia H100 在性能上仍存在较大差距。同时国产芯片公司仍在加快研发推进下一代 AI 芯片产品，并有望在未来对标 Nvidia H100，如寒武纪在研的思元 590、沐曦在研的 MXC500 等。

表 6：国产 AI 芯片性能指标仍与国际顶尖水平存在较大差距

公司	Nvidia	寒武纪	华为	沐曦
产品型号	H100 SXM	思元 370	昇腾 910	曦思 N100
FP32	67TFlop/s	24TFlop/s	/	/
FP16	1979TFlops/s	96TFlop/s	320TFlop/s	80TFlop/s
INT8	3958Top/s	256Top/s	640Top/s	160Top/s

资料来源：各公司公告，源达信息证券研究所

美国对 AI 芯片出口管制，自主可控要求下国产芯片需求迫切。2022 年 10 月 7 日美国商务部工业安全局（BIS）发布《美国商务部对中华人民共和国（PRC）关于先进计算和半导体实施新的出口管制制造》细则，其中管制物项 3A090、4A090 包含高性能 AI 芯片产品，而 Nvidia A100 和 H100 均符合管制要求。在此背景下，Nvidia 推出性能阉割的中国特供版芯片 A800 和 H800。我们认为在国内自主可控大背景下，国内 AI 产业对国产芯片需求迫切，或加大对国产芯片公司支持力度，国产 AI 芯片有望迎来技术进步和市场机遇。

表 7：BIS 禁令限制高性能 AI 芯片向中国出口

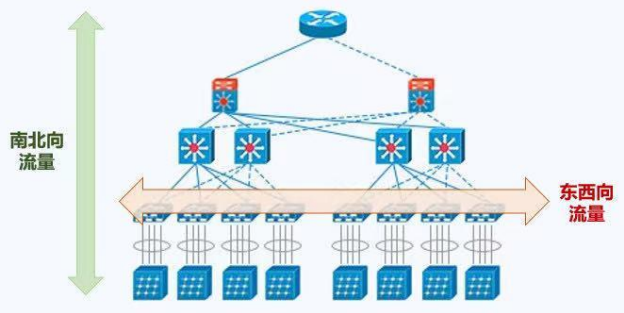
管制物项	管制范围
3A090	1、输入输出（I/O）双向传输速度高于 600GB/s； 2、算力性能与精度指令比特长度乘积超过 4800
4A090	1、含有超过 3A090 技术指标芯片的计算机、电子组件和相关部件

资料来源：美国商务部，源达信息证券研究所

五、算力产业链：光模块快速放量，产品结构向高端升级

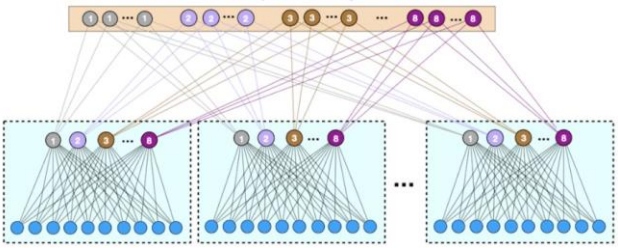
高算力需要与高效传输架构相匹配。AI 大模型通常由多个服务器作为节点，并通过高速网络架构组成集群合作完成模型训练。因此在模型中东西向流量（数据中心服务器间的传输流量）大幅增加，而模型训练过程中南北向流量（客户端与服务器间的传输流量）较少，由于叶脊网络架构相较传统三层架构更适用于东西向流量传输，成为现代数据中心主流网络架构。

图 28：AI 大模型中东西向流量显著增加



资料来源：华为云，源达信息证券研究所

图 29：叶脊网络架构适用于东西向流量传输



资料来源：鹅厂网事，源达信息证券研究所

叶脊网络架构大幅增加对光模块数量需求。由于叶脊网络架构中东西向流量大，因此服务器与交换机相连均需使用光模块，从而大幅增加对光模块数量需求。同时 AI 大模型的高流量对带宽提出更高要求，800G 光模块相较 200G/400G 光模块具有高带宽、功耗低等优点，有望在 AI 大模型网络架构中渗透率提升。

表 8：叶脊网络架构对光模块数量需求大幅提升

架构类型	传统三层架构	改进等三层架构	叶脊网络架构
光模块相对于机柜倍数	8.8	9.2	44/48

资料来源：中际旭创定向增发募集说明书，源达信息证券研究所

我们以 Nvidia DGX H100 网络架构为例。该架构适配 Nvidia H100 GPU，采用叶脊网络架构，分为 1-4 个 SU 单元类型（8 个 GPU 组成一个 H100 服务器节点，32 个服务器节点组成一个 SU 单元）。其中 4-SU 单元架构由 127 个服务器节点组成（其中一个节点用于安装 UFM 网络遥测装置），具有 1016 个 H100 GPU、32 个叶交换机、16 个脊交换机。

表 9：Nvidia DGX H100 架构所需 GPU、交换机数量

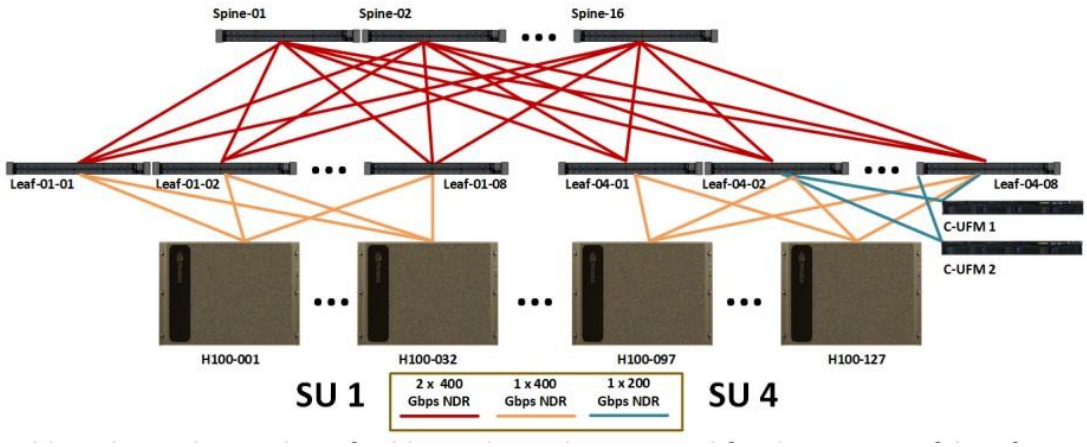
SU Count	Cluster Size # Nodes	Cluster Size # GPUs	Leaf Switch Count	Spine Switch Count	Compute + UFM Node Cable Count	Spine-Leaf Cable Count
1	31[1]	248	8	4	252	256
2	63	504	16	8	508	512
3	95	760	24	16	764	768
4	127	1016	32	16	1020	1024
1. This is a 32 node per SU design, however a DGX Node must be removed to accommodate for UFM connectivity.						

资料来源：Nvidia，源达信息证券研究所

我们以 Nvidia DGX H100 架构为例测算 GPU 与光模块的对应数量。在 4-SU 的 Nvidia DGX H100 架构中，每 32 台服务器节点组成一个 SU 单元，并与 8 台叶交换机相连，因此服务器节点与叶交换机之间共有 1024 个连接（32×8×4）；32 台叶交换机需分别与 16 台脊交换机相连，因此叶交换机与脊交换机之间共有 512 个连接（32×16）；

在 Nvidia DGX H100 的目前方案中，脊-叶连接采用 800G 光模块，需要 1024 个 800G 光模块；叶-服务器连接中，每个服务器节点通过一个 800G 光模块与两台叶交换机向上连接，需要 512 个 800G 光模块（128×4），同时每台叶交换机通过一个 400G 光模块与一个服务器节点连接，需要 1024 个 400G 光模块（128×8）。综上计算得一个 4-SU 单元的 DGX H100 架构需要 1016 个 GPU、1536 个 800G 光模块、1024 个 400G 光模块，**GPU：800G 光模块：400G 光模块的比例约等于 1：1.5：1。**

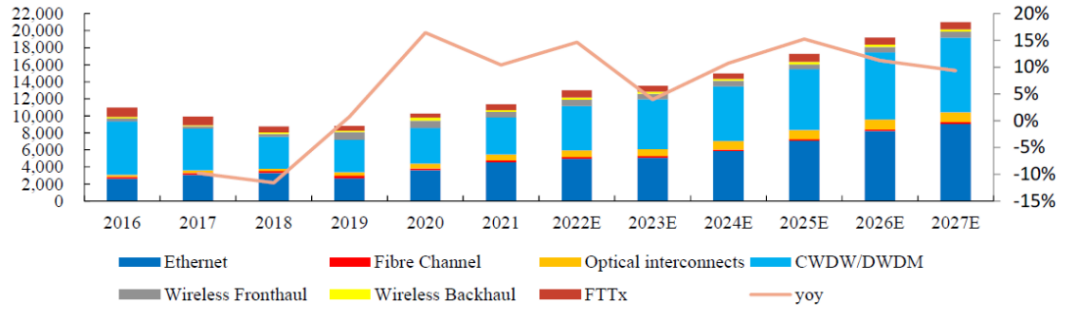
图 30：Nvidia DGX H100 架构示意图



资料来源：Nvidia，源达信息证券研究所

根据 Lightcounting 预测，全球光模块市场规模在 2025 年有望达 170 亿美元，并在 2025-2027 年或将以 CAGR=11%的复合增速增长，2027 年有望突破 200 亿美元。

图 31：全球光模块市场在 2027 年有望突破 200 亿美元



资料来源：Lightcounting，源达信息证券研究所

国产光模块厂商在 2023 年全球光模块企业 TOP10 排名中占据 7 席。TOP10 中国内企业为中际旭创（Innolight）、华为（Huawei）、光迅科技（Accelink）、海信（Hisense）、新易盛（Eoptolink）、华工正源（HGG）、索尔思光电（已被华西股份收购）。而在高端光模块领域，中际旭创已在 2024 年实现 1.6TG 光模块批量出货，并加快对 3.2T 等更高端光模块技术的研发。

表 10：中际旭创在 2023 年全球光模块企业排名中位居第一

2021	2022	2023
Ⅱ-Ⅳ&Innolight	Innolight&Coherent	Innolight Coherent
Huawei（HiSilicon）	Cisco(Acacia)	Huawei（HiSilicon）
Cisco（Acacia）	Huawei（HiSilicon）	Cisco(Acacia)
Hisense	Accelink	Accelink
Broadcom（Avago）	Hisense	Hisense
Eoptolink	Eoptolink	Eoptolink
Accelink	HGG	HGG
Molex	Intel	Source Photonics
Intel	Source Photonics	Marvell

资料来源：Lightcounting，源达信息证券研究所

六、投资建议

1.建议关注

豆包大模型产品力大幅提升，产品矩阵进一步丰富，并大幅降低人工智能大模型使用成本，配合豆包 APP 在终端应用渗透率的提升，有望推动 AI 应用的商业繁荣。伴随 AI 应用需求增长，打开算力高额需求缺口，推动算力基础设施建设。建议关注服务器、液冷设备、芯片和光模块等领域的投资机会：

1）服务器：浪潮信息、中科曙光；

2）液冷设备：英维克；

3）芯片：海光信息；

4）光模块：中际旭创、天孚通信、光迅科技。

2.行业重点公司一致盈利预测

表 11：万得一致盈利预测

公司	代码	归母净利润（亿元）			PE			总市值（亿元）
		2023E	2024E	2025E	2023E	2024E	2025E	
浪潮信息	000977.SZ	22.9	28.5	34.0	34.2	27.5	23.0	781
中科曙光	603019.SH	21.7	26.6	32.0	50.0	40.7	33.9	1083
英维克	002837.SZ	5.4	7.2	9.4	56.1	41.9	32.3	303
海光信息	688041.SH	19.0	27.2	36.5	187.9	131.7	98.1	3579
中际旭创	300308.SZ	53.4	88.0	108.9	27.5	16.7	13.5	1471
天孚通信	300394.SZ	14.4	23.2	30.3	36.9	22.9	17.6	531
光迅科技	002281.SZ	7.8	10.9	13.7	56.7	40.5	32.3	440

资料来源：Wind 一致预期（2024/12/30），源达信息证券研究所

七、风险提示

算力资本开支不及预期；

AI 应用渗透不及预期；

宏观经济环境恶化；

竞争格局恶化。

投资评级说明

行业评级	以报告日后的 6 个月内，证券相对于沪深 300 指数的涨跌幅为标准，投资建议的评级标准为：		
	看	好：	行业指数相对于沪深 300 指数表现＋10%以上
	中	性：	行业指数相对于沪深 300 指数表现 - 10%~＋10%以上
	看	淡：	行业指数相对于沪深 300 指数表现 - 10%以下
公司评级	以报告日后的 6 个月内，行业指数相对于沪深 300 指数的涨跌幅为标准，投资建议的评级标准为：		
	买	入：	相对于恒生沪深 300 指数表现＋20%以上
	增	持：	相对于沪深 300 指数表现＋10%~＋20%
	中	性：	相对于沪深 300 指数表现 - 10%~＋10%之间波动
	减	持：	相对于沪深 300 指数表现 - 10%以下

办公地址

石家庄
河北省石家庄市长安区跃进路 167 号源达办公楼

上海
上海市浦东新区峨山路 91 弄 100 号陆家嘴软件园 2 号楼 701 室

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并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。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。作者所得报酬的任何部分不曾与，不与，也不将与本报告中的具体推荐意见或观点而有直接或间接联系，特此声明。

重要声明

河北源达信息技术股份有限公司具有证券投资咨询业务资格，经营证券业务许可证编号：911301001043661976。

本报告仅限中国大陆地区发行，仅供河北源达信息技术股份有限公司（以下简称：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。本报告的信息均来源于公开资料，本公司对这些信息的准确性和完整性不作任何保证，也不保证所包含信息和建议不发生任何变更。本公司已力求报告内容的客观、公正，但文中的观点、结论和建议仅供参考，不包含作者对证券价格涨跌或市场走势的确定性判断。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，投资者应当对本报告中的信息和意见进行独立评估。

本报告仅反映本公司于发布报告当日的判断，在不同时期，本公司可以发出其他与本报告所载信息不一致及有不同结论的报告；本报告所反映研究人员的不同观点、见解及分析方法，并不代表本公司或其他附属机构的立场。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。

本公司及作者在自身所知情范围内，与本报告中所评价或推荐的证券不存在法律法规要求披露或采取限制、静默措施的利益冲突。