

AI 新范式：云厂商引领+内需为王

2024 年 12 月 29 日

► **Scaling Law 2.0, CSP 的私域数据成为关键。**过去大模型的发展符合 Scaling Law, 然而当下公有数据逐步达到瓶颈, **私域高精度数据或成为 Scaling Law 2.0 的核心要素。**Open AI 前首席科学家 Ilya 在公开演讲中提到, 由于“我们已经达到了数据的峰值”, 当前 AI 模型的预训练方式可能走向终结。想要在特定领域训练出垂直化的“专家大模型”, 数据的精度、准确度等指标更为重要, 私域数据、人工标注的数据可能成为下一阶段大模型发展过程中的核心竞争力, **掌握私域数据的 CSP 厂商将在大模型厂商的下一轮竞赛中更具优势。**

► **CSP 异军突起, 百舸争流的算力竞争时代开启。**当下 AI 大模型日益增长的算力需求对 CSP 厂商资本开支提出更高要求, 考虑到成本优势以及更好的灵活性, CSP 自研加速卡将成为未来 AI 芯片增量最核心的来源, 英伟达在加速卡领域的市占率有望逐步被 CSP 厂商替代。与此同时, **CSP 自研算力也带来了算力产业链的全新变革, 全新的 AEC、PCB、散热、电源产业链冉冉升起, CSP 合作伙伴将在算力的下一个环节中更为受益。**

► **端侧：豆包出圈, 互联网巨头入局 AI 终端。**AI 终端的发展经历了由云到端、由 ToB 到 ToC 的 5 个阶段, **当前我们处于以字节豆包为代表的互联网巨头引领的第五阶段。**早期 AI 终端更多是品牌厂商引领, 而当前互联网巨头亲自下场开发硬件, 有望在 AI 终端产业链中发挥更加重要的作用。我们看好 AI+智能终端的趋势, AI 将重构电子产业的成长, 加速硬件的智能化、伴侣化趋势。**无论是手机、PC、AIOT、可穿戴设备、汽车电子, 都有重估的潜力。**

► **投资建议：我们认为, 伴随着公开数据预训练 scaling law 的结束, AI 产业叙事开始向云厂商合作伙伴转移。**从数据到模型, 从训练到推理, 从云到端, CSP 厂商全面布局, 形成了完美的商业闭环。当下, 无论是海外的谷歌、亚马逊, 还是国内的字节、腾讯, 云厂商巨头们开始接力, 引领 AI 产业的下一棒。具体到投资方向, PCB、铜缆、温控、电源等产业链, 是国内企业深耕多年, 具备优势的环节。伴随着本土云厂商的大力扩产, 内需为王的时代也将来临。相比过去, 25 年的 AI 产业投资将更重视合作建立、订单落地以及业绩兑现, 投资回报也会更为稳健。**建议关注：1、云端算力：1) ASIC：寒武纪、海光信息、中兴通讯；2) 服务器：浪潮信息、工业富联、华勤技术、联想集团；3) AEC：新易盛、博创科技、瑞可达、兆龙互联、立讯精密；4) 铜连接：沃尔核材、精达股份；5) PCB：生益电子、广合科技、深南电路、威尔高；6) 散热：申菱环境、英维克、高澜股份；7) 电源：麦格米特、欧陆通、泰嘉股份。2、端侧硬件：1) 品牌：小米集团、漫步者、亿道信息等；2) 代工：国光电器、歌尔股份、天键股份、佳禾智能等；3) 数字芯片：乐鑫科技、恒玄科技、星辰科技、富瀚微、中科蓝讯、炬芯科技、全志科技等；4) 存储芯片：兆易创新、普冉股份；5) 渠道配镜：博士眼镜、明月镜片等。**

► **风险提示：**大模型发展不及预期, CSP 自研算力不及预期, AI 终端销量不及预期。

重点公司盈利预测、估值与评级

代码	简称	股价 (元)	EPS (元)			PE (倍)			评级
			2024E	2025E	2026E	2024E	2025E	2026E	
688041.SH	海光信息	155.05	0.84	1.12	1.46	185	138	106	推荐
688800.SH	瑞可达	56.19	1.04	1.38	1.84	54	41	31	推荐
002851.SZ	麦格米特	60.72	1.13	1.53	1.95	54	40	31	/
300870.SZ	欧陆通	111.99	2.11	2.87	3.91	53	39	29	/
300499.SZ	高澜股份	22.16	0.10	0.27	0.43	218	83	52	/
301018.SZ	申菱环境	38.98	0.77	1.02	1.24	51	38	31	/
002130.SZ	沃尔核材	29.22	0.73	0.95	1.14	40	31	26	/
600577.SH	精达股份	8.02	0.26	0.34	0.41	31	23	20	/
688629.SH	华丰科技	38.89	0.20	0.46	0.63	196	84	62	/

资料来源：, 民生证券研究院预测；

(注：股价为 2024 年 12 月 27 日收盘价；未覆盖公司数据采用 wind 一致预期)

推荐

维持评级



分析师 方竞

执业证书：S0100521120004

邮箱：fangjing@mszq.com

分析师 宋晓东

执业证书：S0100523110001

邮箱：songxiaodong@mszq.com

研究助理 李伯语

执业证书：S0100123040030

邮箱：liboyu@mszq.com

相关研究

1. 半导体行业 2025 年度投资策略：如鱼跃渊，升腾化龙-2024/12/25
2. EDA 和 IP 行业专题：半导体产业基石，国产替代打破垄断格局-2024/12/25
3. 半导体行业点评：国产 DDR5 突破，看好 D RAM 产业链-2024/12/23
4. 电子行业动态：豆包出圈，解析字节的 AI 终端布局-2024/12/18
5. 电子行业点评：11 月手机补贴落地，12 月 iOS18.2 将至-2024/12/06

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

目录

1 Scaling Law 2.0, CSP 的私域数据成为关键	3
1.1 关于 Scaling Law 的争议, 从数据规模到数据精度	3
1.2 CSP 的三大利器: 私域数据、推理需求、从云到端	5
2 CSP 异军突起, 百舸争流的算力竞争时代开启	9
2.1 从外采到自研, CSP 的算力升级之路	9
2.2 CSP 算力供应链新变革	16
3 端侧: 豆包出圈, 互联网巨头入局 AI 终端	28
3.1 字节豆包先行, 加速端侧落地	28
3.2 AI 终端空间广阔, SoC 是影响体验的核心硬件	32
4 投资建议	36
4.1 行业投资建议	36
4.2 相关公司梳理	36
5 风险提示	40
插图目录	41
表格目录	41

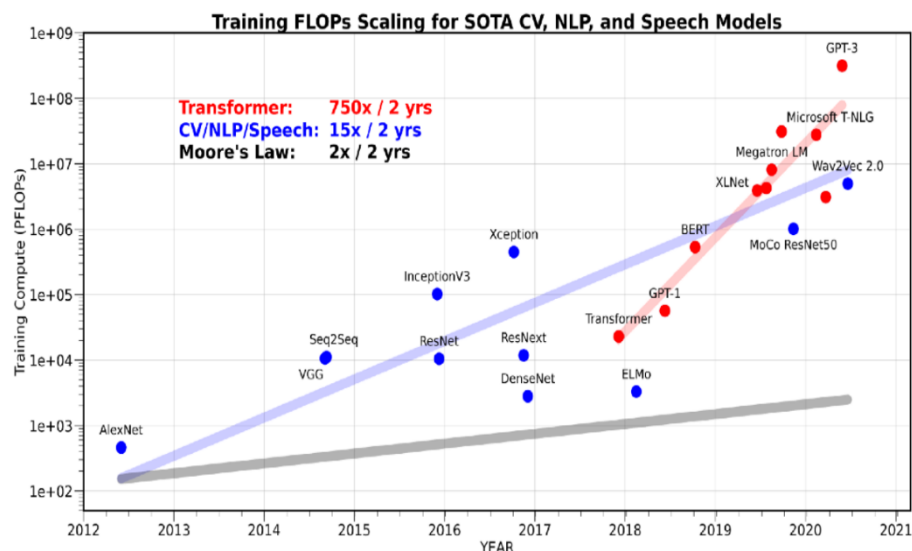
1 Scaling Law 2.0，CSP 的私域数据成为关键

1.1 关于 Scaling Law 的争议，从数据规模到数据精度

1.1.1 大模型的摩尔定律，算力需求指数级增长

Scaling Law 是 AI 产业发展的通用规律，在 Scaling Law 下，大模型对算力的需求以每年 10 倍左右的速度增长，甚至超过了摩尔定律下半导体晶体管密度的增长速度。AI 大模型的算力需求在过去几年呈现快速增长的态势，Transformer 算力需求在 2 年内增长 750 倍，平均每年以接近 10 倍的速度增长。以 Open AI 的 GPT 为例，GPT 1 在 2018 年推出，参数量级为 1 亿个，Open AI 下一代推出的 GPT 5 参数量级预计达到 10 万亿。

图1：AI 大模型对算力的需求超过摩尔定律

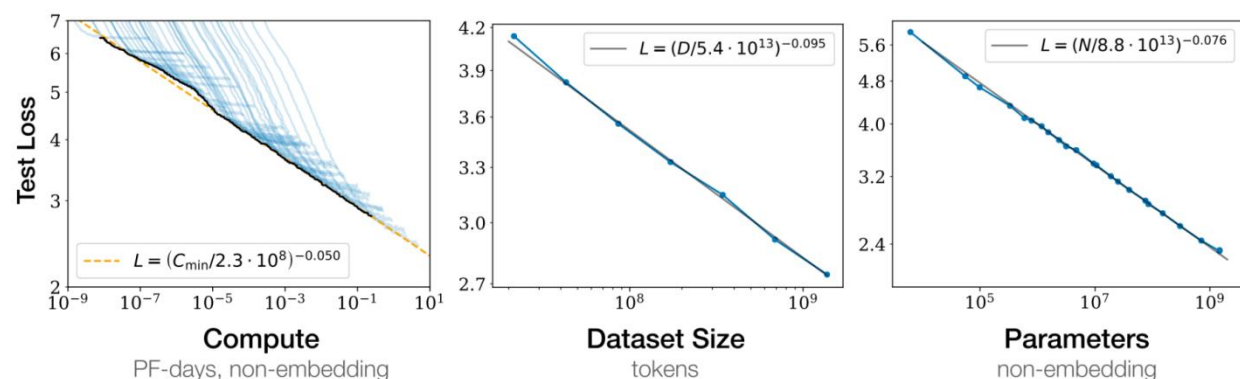


资料来源：CSDN，民生证券研究院

1.1.2 数据成为瓶颈，Scaling Law 放缓

大模型的 Scaling Law 表明，计算量、数据量、参数规模三个因素的增长能够不断提升大模型的性能。在任意其他两个指标不受限制的情况下，大模型的性能和另一个因素都呈现幂律关系，在大模型过去的发展过程中，算力、数据量、参数规模三个指标均没有达到上限，Scaling Law 仍然在发挥作用，大模型的性能也在持续改善。

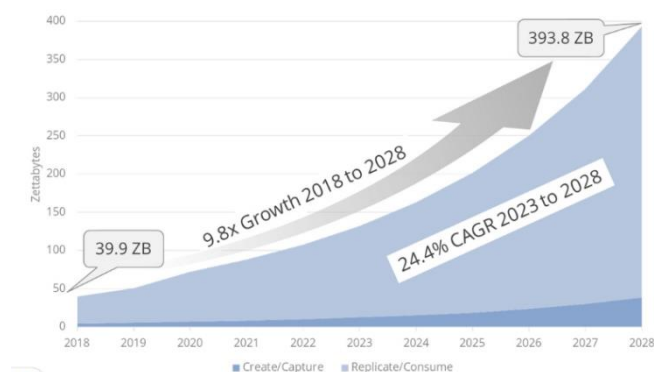
图2: Scaling Law 的三要素: 算力、数据量、参数规模



资料来源: Elias Z. Wang 《Scaling Laws for Neural Language Models》, 民生证券研究院

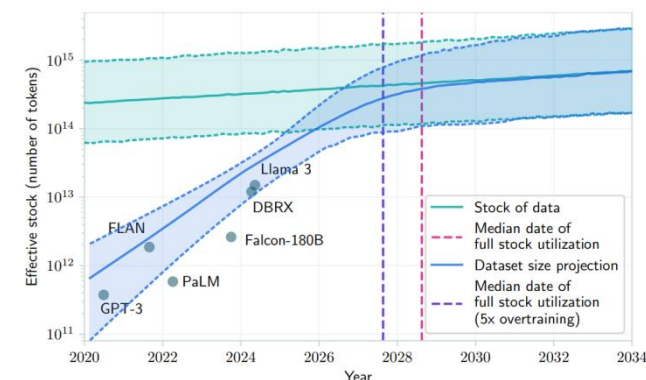
然而公开数据量的有限性制约了 Scaling Law 进一步发挥作用。据 IDC, 2018 年全球数据总量为 39.9ZB, 预计到 2028 年, 全球数据总量将达到 393.8ZB, CAGR 增速为 25.7%, 该增速远远低于 Scaling Law 下大模型参数和算力需求每年 10 倍左右的增长速度。Pablo Villalobos 等人的研究表明, 在 2028 年左右, 大模型能够获得的数据量级将达到上限, 受限数据量, Scaling Law 将会放缓。实际上, 由于大模型自 2022 年底以来的加速发展, 数据量可能在 2028 年以前就会达到天花板, 从而限制 Scaling Law 发挥作用。

图3: 2018-2028 年全球数据量



资料来源: IDC, 民生证券研究院

图4: 大模型 Scaling Law 将在 2028 年左右开始显著放缓



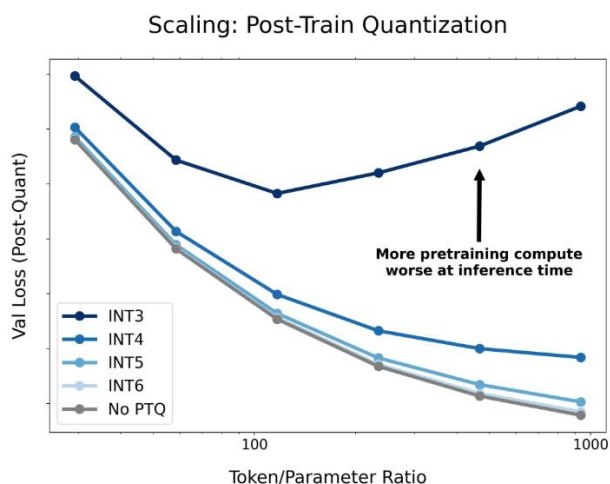
资料来源: 《Will we run out of data? Limits of LLM scaling based on human-generated data》Pablo Villalobos 等, 民生证券研究院

1.1.3 Scaling Law 2.0, 高精度私域数据的强化学习

当下传统的 Scaling Law 受限数据量, 私域高精度数据或成为 Scaling Law 2.0 的核心要素。12 月 15 日, 在 NeurIPS 大会上, Open AI 前首席科学家 Ilya 在公开演讲中提到, 由于目前“我们已经达到了数据的峰值, 未来不会再有更多的数据”, 当前 AI 模型的预训练方式可能走向终结。Ilya 的发言认为当前传统的 Scaling Law 即将失效, 新的 Scaling Law, 即在特定领域的强化学习将发挥更

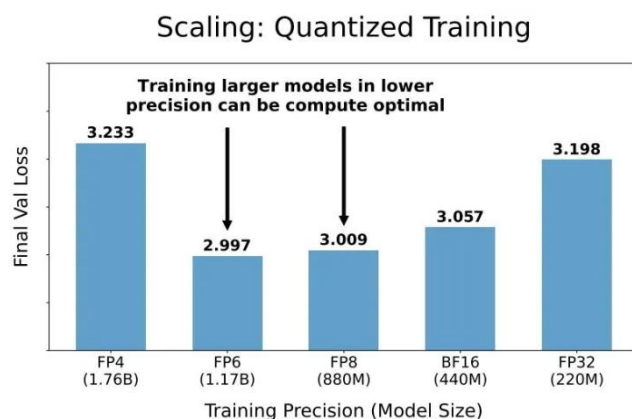
重要的作用。想要在特定领域训练出垂直化的“专家大模型”，数据的数量不再成为衡量数据好坏的唯一标准，数据的精度、准确度等指标更为重要，私域数据、人工标注的数据可能成为下一阶段大模型发展过程中的核心竞争力。

图5：低精度的训练数据增多可能反而对模型性能造成损害



资料来源：《Scaling Laws for Precision》，民生证券研究院

图6：使用更高精度的数据将减小因数据质量不佳而对模型性能造成的损害



资料来源：《Scaling Laws for Precision》，民生证券研究院

1.2 CSP 的三大利器：私域数据、推理需求、从云到端

1.2.1 CSP：掌握私域数据，延续 Scaling Law

私域数据成为延续 Scaling Law 的关键。随着大模型规模及训练集的扩大，可用于训练的互联网公开数据逐渐稀少，基于公开数据的 Scaling Law 逐步走到尽头；拥有庞大用户群体的互联网巨头，海外以 AWS、Meta、谷歌、微软代表，国内以字节、阿里、腾讯、百度等为代表，掌握了用户在使用过程中产生的大量私域数据，如 AWS 依托其购物平台，积攒了大量用户偏好及使用习惯数据，以及大量的商品文字/图片数据；字节依托头条及抖音平台，积累了海量的资讯数据及视频数据，可以为模型的训练提供优质的语料，有力解决高质量数据匮乏导致的“Scaling Law 失效”问题。

图7：豆包大模型使用量快速增长



资料来源：字节跳动发布会，民生证券研究院

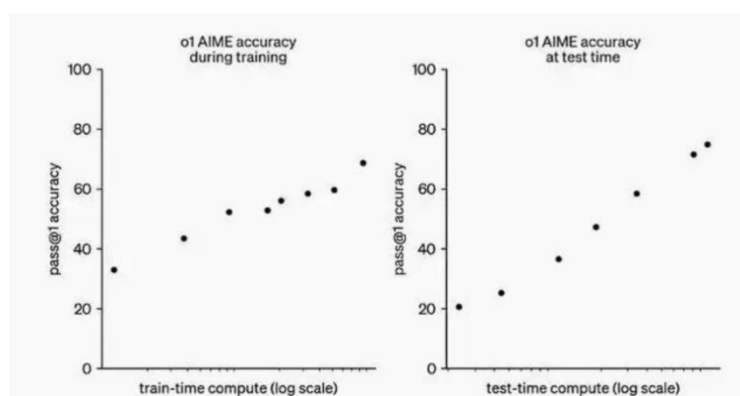
1.2.2 CSP：优质流量接口赋能推理侧，具备推理成本优势

算力需求会加速从训练侧向推理侧倾斜，推理有望接力训练，成为下一阶段算力需求的主要驱动力。互联网巨头旗下亦拥有高活跃应用可以帮助引流，可以依托自身平台流量推广自家大模型，从而获取客户、积累数据；可以将大模型进一步绑定手中热门应用，如可以让自家大模型成为自家短视频平台辅助剪辑、修图的工具。**CSP 优质的流量接口可赋能推理侧，实现从训练模型到用户使用大模型进行推理的正向循环，加速模型商业化落地。**

随着端侧 AI 的放量及 AI 应用的落地，豆包、ChatGPT 等 AI 应用快速发展，AI 推理计算需求将快速提升，巴克莱的报告预计 AI 推理计算需求将占通用人工智能总计算需求的 70%以上，**推理计算的需求甚至可以超过训练计算需求，达到后者的 4.5 倍。**

OpenAI 上线的 o1 模型也更加侧重于推理测能力。o1 新模型除预训练阶段外，还通过大量计算反复强化学习，并且在推理阶段增加思考时间，获得能力的提升，未来模型通过大量强化学习或者延长思考时间以获得更准确的答案或成为可能。英伟达高级科学家 Jim Fan 认为 o1 模型的推出意味着模型开始呈现出推理侧的 Scaling law，双曲线的共同增长，有望突破大模型能力的提升瓶颈。开发者访问 o1 的成本更高，百万 token 输入/输出收费为 GPT-4o 的 3/4 倍；**复杂推理明显拉动推理侧算力需求。**

图8：推理对模型重要性提升



资料来源：OpenAI，民生证券研究院

鉴于推理所需的芯片算力显著低于训练，尽管 AWS、谷歌为代表的 CSP 自研 ASIC 起步较晚，整体性能低于英伟达的 GPU，但由于其较高的性价比，以及定制化的开发更加匹配自家模型的推理，**云厂商可以通过 ASIC 降低推理成本，扩大自身优势**。博通于业绩会透露，目前正在与三个非常大型的客户开发 AI 芯片，预计明年公司 AI 芯片的市场规模为 150 亿-200 亿美元。英伟达 GPU 目前在推理市场中市占率约 80%，但随着大型科技公司定制化 ASIC 芯片不断涌现，这一比例有望在 2028 年下降至 50%左右。

图9：谷歌 TPU



资料来源：谷歌，集微网，民生证券研究院

1.2.3 CSP：大模型赋能终端的重要参与者

硬件为大模型落地最重要的载体，从云到端为大模型落地的必经之路。 复盘

AI 发展之路，从大模型推出以来，AI PC、AI+MR、AI 手机、AI 眼镜等终端陆续落地。我们认为 AI 终端的定价=硬件成本+AI 体验，大模型体验的优劣决定了用户的付费意愿，而互联网巨头和第三方科技公司深度参与模型的开发，部分模型能力较强的厂商如字节、百度甚至亲自下场开发硬件，因此我们认为互联网巨头有望在 AI 终端产业链中发挥更重要的作用。

图10：AI 发展历程复盘



资料来源：联想、苹果、Open AI 官网等，民生证券研究院

1.2.4 结论：AI 发展的话语权将转换至 CSP

我们认为，随着公开数据 Scaling Law 的逐步终结，掌握私域数据的云厂商更有希望延续大模型训练的 Scaling Law；随着 AI 应用的逐步落地，AI 将由训练端逐步转向推理端，云厂商自研 ASIC 将比通用 GPU 更具性价比，且自身具备优质的流量接口可赋能推理侧；硬件是大模型落地最重要的硬件载体，而 CSP 将作为大模型重要提供方，将全面赋能 AIPC、AI 手机、AI 眼镜、机器人等，CSP+电子品牌（苹果、华为、联想等）的商业模式将更为清晰。

从公开数据到私域数据，从训练到推理，从云到端，三大利器全部掌握在 CSP 手中，形成训练-推理-商业化落地的完美闭环，我们认为未来 CSP 的话语权将进一步加重，AI 发展的话语权将由以英伟达为代表的算力公司，以及以 Open AI 为代表的大模型公司，交棒至各大 CSP 巨头。

2 CSP 异军突起，百舸争流的算力竞争时代开启

2.1 从外采到自研，CSP 的算力升级之路

2.1.1 资本开支和云收入，相辅相成

云商算力需求仍维持高增，持续增长的资本开支成为北美云商算力的主要支撑。CY3Q24 北美四大云商 CY3Q24 合计资本开支为 598.14 亿美元，同比增长 61.7%，环比增长 13.2%；Bloomberg 一致预期 CY2024 北美四大云商资本开支合计为 2226 亿美元，同比增长 51.0%。其中：

1) 微软：CY3Q24 资本开支 149.23 亿美元，qoq+7.6%，yoy+50.5%，此前公司指引 Q3 资本开支环比增加，实际数据基本符合预期，公司指引下一季度资本开支环比增长，但伴随需求增长，资本开支增速将放缓；

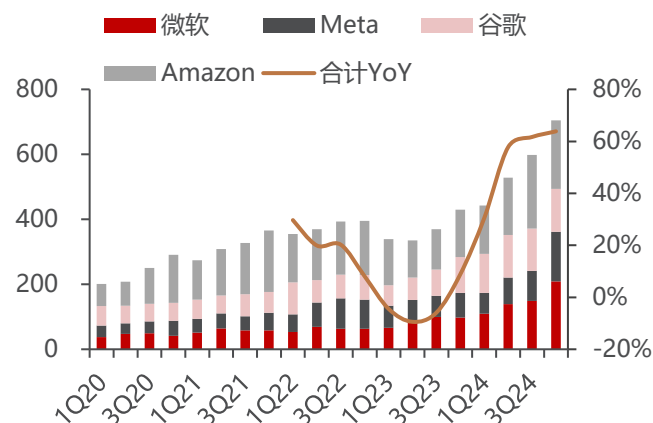
2) Meta：CY3Q24 资本开支 92.10 亿美元，qoq+12.7%，yoy+40.8%，公司将全年资本开支预期范围从上一季度的 370-400 亿美元上修至 380-400 亿美元，符合 391 亿美元的 Bloomberg 一致预期，关于 2025 年，公司预计 2025 年资本支出将继续大幅增长；

3) 谷歌：CY3Q24 资本开支 130.61 亿美元，qoq-0.9%，yoy+62.10%，此前公司指引 Q3 资本开支保持或高于 120 亿美元，实际数据略超预期，据公司指引我们测算公司 2024 年资本开支预计 500 亿美元以上，2025 年保持适度增长，与此前指引一致；

4) 亚马逊：CY3Q24 资本开支 226.20 亿美元，qoq+28.4%，yoy+81.3%，显著超 Bloomberg 一致预期，公司对 2024 年资本开支计划为 750 亿美元，预计 2025 年将进一步提升。

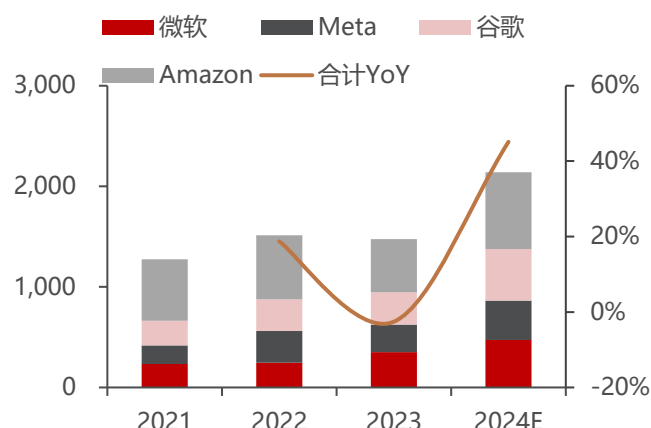
我们认为，当前海外算力产业链的核心矛盾为云商资本开支和算力需求的增长。从北美四大云厂商对资本开支最新的反馈来看，算力需求依旧旺盛。另一方面，云商正在通过：1) AI 带来 Opex 方面的降本增效；2) 延长云基础设施的折旧年限等方式提升资本开支潜在空间，算力需求有望持续高增。

图11: 1Q20-4Q24 北美云商资本开支 (亿美元)



资料来源: Bloomberg, 民生证券研究院
注: 4Q24 数据为 Bloomberg 一致预期

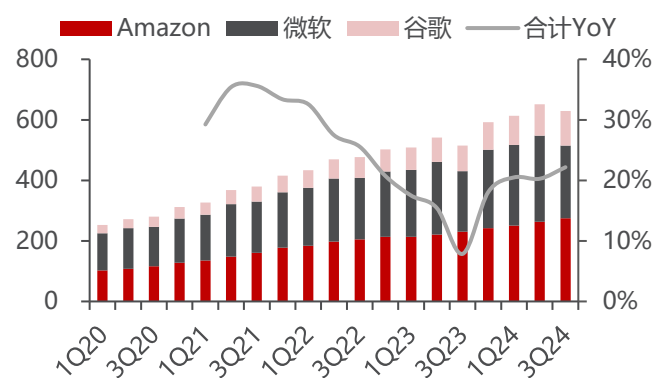
图12: 2021-2024E 北美云商资本开支 (亿美元)



资料来源: Bloomberg, 民生证券研究院

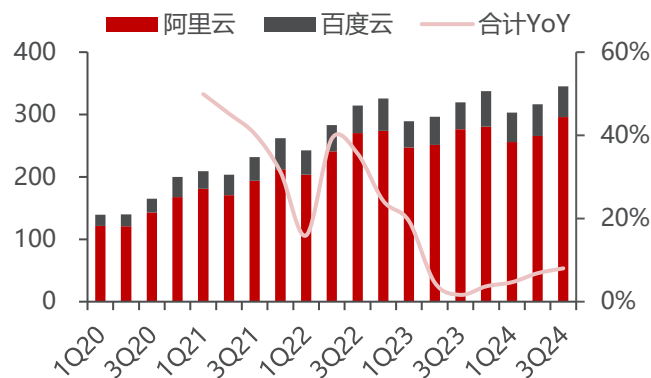
云商云业务收入持续高增, 为算力基础设施扩张提供主要动力。北美云商方面, CY1Q23-CY3Q24 亚马逊、微软和谷歌三大云商合计云业务收入分别为 508.89/541.64/514.83/592.76/613.19/651.43/628.97 亿美元, 同比增长 17.5%/15.4%/7.9%/18.1%/20.5%/20.3%/22.2%, 生成式 AI 的发展使得用户对云的需求持续高涨; 国内云商方面, CY1Q23-CY3Q24 阿里和百度两大云商合计云业务收入分别为 289.42/296.23/319.48/337.66/302.95/316.49/345.10 亿元, 同比增长 19.4%/4.6%/1.6%/3.7%/4.7%/6.8%/8.0%, 增速有所放缓, 但国内云行业正向 AI 计算加速转变, 未来上升趋势依然可观。

图13: 北美云商云业务收入 (亿美元) 及同比增速



资料来源: Bloomberg, 民生证券研究院

图14: 国内云商云业务收入 (亿美元) 及同比增速



资料来源: Bloomberg, 民生证券研究院

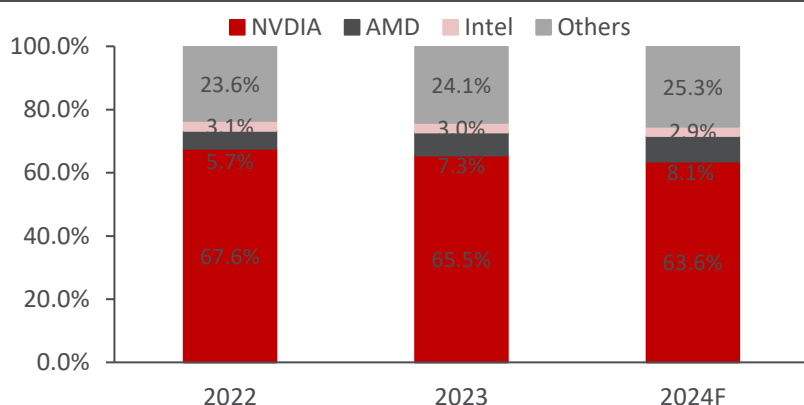
2.1.2 CSP 加速入局, 挑战英伟达垄断

目前英伟达在 AI 加速卡领域仍然占据主要供应商的低位, 但其他商业公司以及云商自研的比例正在逐步提升。AI 加速卡的竞争格局中主要有三类“玩家”:
1) 全球龙头英伟达; 2) 其他商业 AI 加速卡公司, 如 AMD、Intel、昇腾等; 3) 云厂商自研加速卡。据 Trendforce 统计, 2022 年全球 AI 芯片市占率来看, 英伟

达独占 67.6%，预计到 2024 年英伟达市占率将下降到 63.6%，而 AMD 以及其他云商自研加速卡比例有望提升。

我们认为,云厂商自研加速卡将成为未来 AI 芯片增量最核心的来源。一方面,云商自研加速卡在成本方面显著优于向英伟达等商业公司外采,3Q24 英伟达毛利率已达到 74.4%，采用自研加速卡的方式,将帮助云商在有限的资本开支下获得更多的 AI 算力。另一方面,云商自研 ASIC 更加灵活,云厂商可以根据自身的模型训练和推理需求,进行 AI 芯片和服务器架构的设计,从而实现更好的训练和推理效果。伴随着云厂商自研 ASIC 产品的逐步成熟,未来云商在 AI 算力的布局中自研的比例有望逐步提升。

图15：2022-2024 年全球 AI 芯片竞争格局



资料来源：Trendforce，民生证券研究院

谷歌采取自研加速卡为主,同时采购部分英伟达加速卡的策略。谷歌研发 TPU 的时间始于 2013 年,相较于其他云厂商有近 10 年的时间优势。由于谷歌在加速卡领域布局早,产品完善度高,谷歌或为 2024 年北美四大云厂商中采购英伟达加速卡最少的厂商。2023 年 12 月,谷歌推出面向云端的 AI 加速卡 TPU v5p,相较于 TPU V4,TPU v5p 提供了二倍的浮点运算能力和三倍内存带宽提升;集群方面,TPU v5p pod 由 8960 颗芯片组成,使用最高带宽的芯片间连接(每芯片 4,800 Gbps)进行互连;从训练效果来看,相较于上一代产品,TPU v5p 训练大型 LLM 的速度提升了 2.8 倍。

Meta 自 2021 年以来便将企业发展的重点放在元宇宙和 AI 领域,并且修改了公司名称。2023 年,Meta 宣布自研 MTIA v1 芯片。2024 年 4 月,Meta 发布最新版本 MTIA v2 加速卡,新一代 MTIA 加速卡在算力、内存容量、内存带宽等方面更加平衡,采用台积电 5nm 工艺制造,Int 8 稀疏算力可以达到 708TOPS, HBM 内存容量达到 128GB。目前 Meta 仍主要采购英伟达等厂商的加速卡用于 Llama 等模型的训练,后续 Meta 有望采用自研加速卡对大语言模型进行训练。

微软 Azure 的企业数量已经达到 25 万家,是目前采购英伟达加速卡最为激进的云厂商,但考虑到采购英伟达加速卡的高昂成本,微软也宣布了自研加速卡的计划。微软的 Maia 100 在 2023 年推出,专为 Azure 云服务设计。Maia 100 采

用台积电 5nm 工艺，单芯片拥有 1050 亿个晶体管，FP 8 算力可以达到 1600TFLOPS，同时支持 FP 4 运算，算力达到 3200TFLOPS，是目前厂商自研加速卡中算力最强大的产品。除了微软自身以外，OpenAI 也在尝试采用微软的加速卡，强大的下游客户支持有望为微软自研加速卡的进步带来重要动力。

亚马逊同样在自研加速卡方面加大投入，并且已经完善了训练和推理的两方面布局。2023 年，亚马逊推出了用于训练的 Trainium2 加速卡，以及用于推理的 Graviton4 加速卡，补全了亚马逊在训练和推理领域加速卡的布局。亚马逊的 Trainium2 加速卡 Int 8 算力达到 861TOPS，相较上一代产品性能提升 4 倍，在云厂商自研加速卡中表现优秀。同时公司的产品可以在新一代 EC2 UltraClusters 中扩展多达 10 万颗 Trainium2 加速卡，与 Amazon EFA PB 级网络互联，提供高达 65 EFlops 算力，为云端大模型的训练和推理提供强大的算力保障。

表1：英伟达客户在加速卡领域的布局情况

厂商	大类	型号	发布时间	制程nm	峰值算力 TOPS/TFLOPS			内存信息		互联带宽GB/s
					INT8/FP8	BF16/FP16	TF32/FP32	类型	容量	
					Dense/Sparse	Dense/Sparse	Dense/Sparse		GB	
谷歌	训练	TPUv5E	2023	-	394	197	-	HBM2	16	400
		TPUv5P	2023	-	918	459	-	HBM2	95	800
Meta	推理	MTIA v2	2024	5	354/708	177/354	2.76	-	128	-
微软	训练	Maia 100	2023	5	1600	800	-	HBM3	64	1200
亚马逊	训练	Trainium2	2023	4	861	431	215	-	96	-
	推理	Graviton4	2023	-	-	-	-	GDDR5	-	-
TESLA	训练	D1	2021	7	362	362	22.6	-	32	-

资料来源：各公司官网，Semianalysis 等，民生证券研究院

注：未标注的数据为没有在公开渠道披露的信息

2.1.3 AWS Trainum 芯片详解

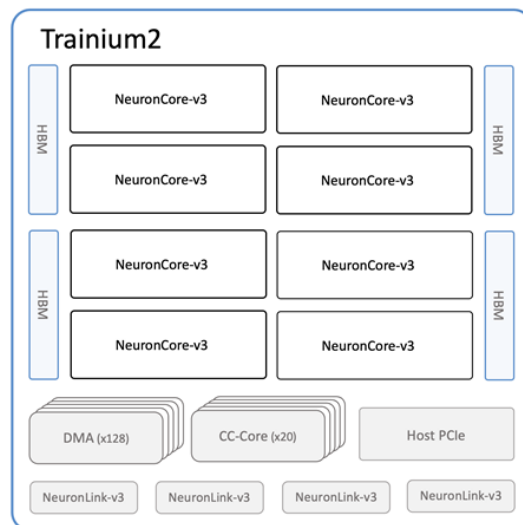
Trainium2 芯片由少量大型 NeuronCore 组成，与 GPU 大量小张量核心不同，该设计更适合生成式 AI 任务。以 NeuronCore 为基础计算单元，在每个 NeuronCore 内配备张量、矢量、标量及 GPSIMD 引擎，分别承担矩阵乘法运算、向量操作、元素级操作及自定义操作，各引擎协同工作提升计算效率。

图16: Trainium2 芯片



资料来源: AWS, 民生证券研究院

图17: Trainium2 内部组件



资料来源: SemiAnalysis, 民生证券研究院

Trainium2 由两个计算芯片组和四个 HBM3e 内存堆栈封装而成, 每个计算芯片经由 CoWoS-S/R 封装与其两个相邻的 HBM3e 进行通信, 芯片的两个部分通过 ABF 基板互连。单芯片具备 650TFLOP/s 的密集 BF16 性能, 1300TFLOPS/s 的密集 FP8 性能, 搭载 96GByte 的 HBM3e 内存容量, 单芯片功耗约 500W。

表2: Trainum2 芯片参数

Trainum2 参数	
	Trn2
Theoretical BF16 Dense TFLOP/s/chip	650
Theoretical FP8 Dense TFLOP/s/chip	1300
HBM Capacity(GByte/chip)	96
HBM Bandwidth(GByte/s/chip)	3200
Arithmetic Intensity(BF16 FLOP per Byte)	203.125
FLOP per HBM Capacity(BF16 FLOP per Byte)	6770.83
Structured Sparsity Support	2:4, 4:8, 4:16
Chip Power(Watts)	500
Cooling Technology	Air Cooled

资料来源: SemiAnalysis, 民生证券研究院

亚马逊的下一代 Trainium3 产品有望实现性能翻倍, 预计将在 25 年底问世。

亚马逊于 12 月 3 日的 re:Invent 2024 大会透露, 首批基于 Trainium3 的实例预计将于 2025 年底上市, 将助力 AWS 及其合作伙伴在 AI 前沿领域取得突破, 满足各行业对 AI 算力增长需求, 推动自然语言处理、图像识别、药物研发等领域创新发展。

Trainium3 将成为首款采用 3nm 工艺节点制造的 AWS 芯片, 在性能、能效和密度上树立了新标杆。更小制程工艺使晶体管尺寸缩小、集成度提高, 降低功耗同时提升运行频率与处理能力, 减少芯片发热, 提升稳定性与可靠性, 为大规模 AI

计算集群节能增效。

Trainium3 芯片能效有望提高 40%、性能有望翻倍提升。更高能效比让单位功耗下计算性能大幅提升，翻倍性能增长使芯片处理能力更强，加速模型训练与推理，支持更大规模复杂模型训练，缩短研发周期、提升迭代效率，增强 AWS 在 AI 芯片领域竞争力。

搭载 Trainium3 的 UltraServer 性能预计将比 Trn2 UltraServer 高出 4 倍。这一性能提升将使得设备在处理复杂 AI 任务时更加高效。对于需要实时处理大量数据的应用场景，如实时语音识别、图像生成等，性能的提升也将带来更流畅的用户体验。

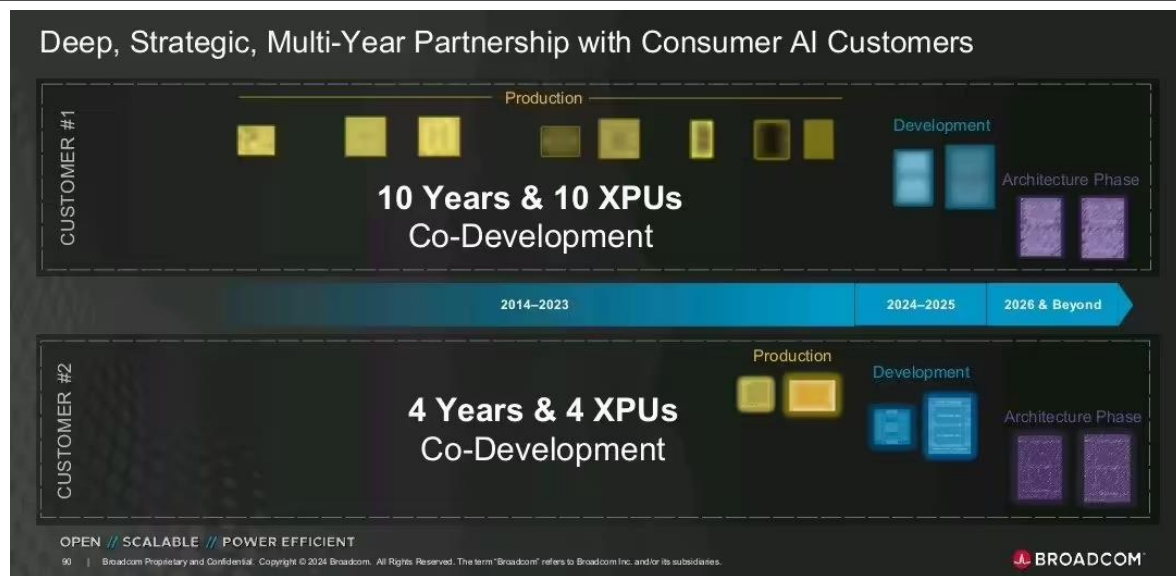
2.1.4 Design Service 成为 CSP 最重要的合作伙伴

相比英伟达的 GPGPU，ASIC 在特定任务场景下，具有高性能、低功耗、成本效益、高保密性和安全性的优势，成为了云厂商提高计算效率，节省算力成本的重要手段。随着 CSP 的话语权不断加强，诸如 AWS、谷歌等对算力的需求不断增长，催生了定制 AI 芯片的需求；**博通、Marvell、Alchip 等可提供先进的 AI 芯片定制化设计服务,成为 CSP 最重要的合作伙伴。**

据 Marvell 预测，2023 年 ASIC 占数据中心加速计算芯片的 16%，规模约为 66 亿美元；随着 AI 计算需求的增长，ASIC 占比有望提升至 25%，预计 2028 年数据中心 ASIC 市场规模将提升至 429 亿美元，CAGR 为 45.4%；博通占据 55% 以上的市场份额，Marvell 市场份额亦接近 15%，成为两个最重要的“玩家”。

博通：谷歌自研 AI 芯片 TPU 的主要供应商，与谷歌 TPU 项目合作约 10 年，客户关系及设计能力积累深厚。公司已与谷歌合作完成其第六代 TPU（Trillium）设计，目前正在推进第七代设计，预计将于 26/27 年推出。受益于 AI 的高速发展，23 年谷歌向博通支付的 TPU 费用达 20 亿美金，24 年有望增长至 70 亿美金。除谷歌外，博通亦为 Meta 的 AI 芯片 MTIA 提供设计服务，亦与苹果、字节跳动、OpenAI 等巨头合作开发芯片。博通 CEO Hock Tan 指出，到 2027 年，博通的客户将在 AI 芯片的集群中部署多达 100 万个芯片，博通的 AI 芯片可能会为公司带来数百亿美元的年收入增长。

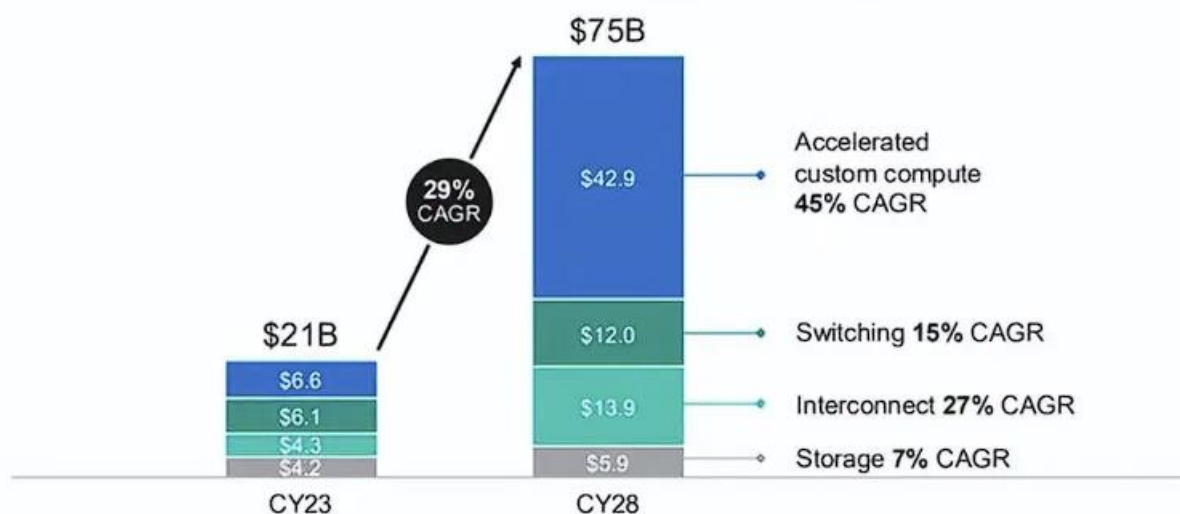
图18: 博通两大核心客户 AI 芯片战略规划



资料来源: IT 之家, 民生证券研究院

Marvell: 亚马逊自研 AI 芯片 Trainium 的主要供应商, 公司与 AWS 签订 5 年合作协议, 帮助亚马逊设计 AI 芯片; 2025 年第三财季 Marvell 面向数据中心的销售额同比增长了近一倍, 达到 11 亿美元, 公司预计本财年数据中心部门将占总收入的 72%, 同比增长 32pts。除 AWS 外, Marvell 亦为微软提供定制的 AI 芯设计服务。Marvell 估计至 28 年, 其计算、高速网络、以太网和存储产品组合将带来 750 亿美元的市场机会, 核心增长动力来自定制计算 (45% 年复合增长率) 和高速数据中心互连 (27% 年复合增长率)。

图19: 博通两大核心客户 AI 芯片战略规划



资料来源: 650 Group, Marvell Investor AI Day, 民生证券研究院

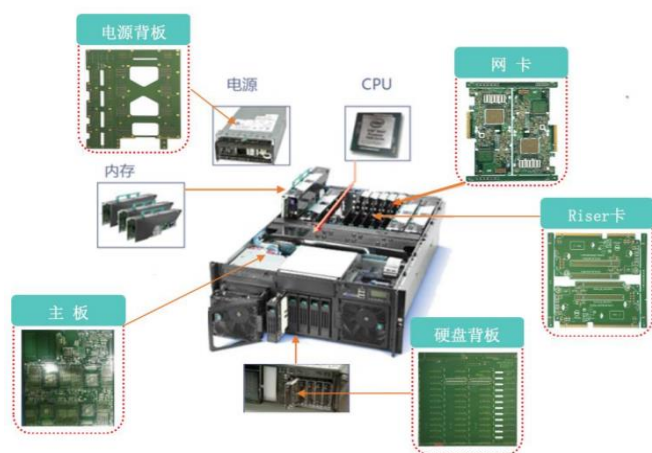
2.2 CSP 算力供应链新变革

2.2.1 拥抱巨头，PCB 供应商深度受益

PCB（印刷电路板，Printed Circuit Board）承担了各电子元器件之间信号互联的功能；AI 服务器中，由于 AI 算力芯片计算速度快、数据吞吐量大，对信号传输有着更高的要求，PCB 的用量、材质及等级亦有提升，价值量相比普通服务器明显提升。

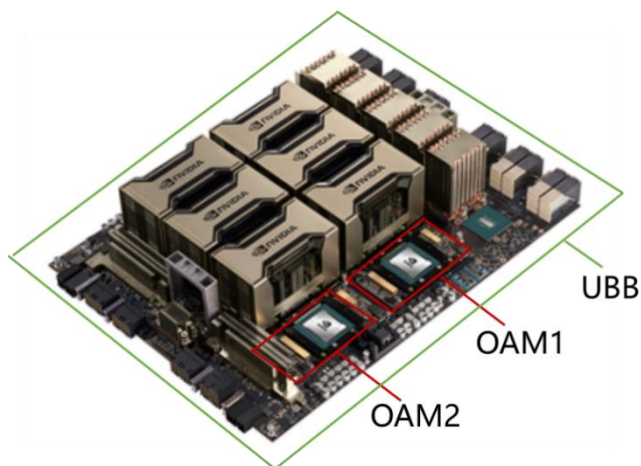
传统的 AI 训练服务器以 8 卡机的形式为主，以英伟达上一代 DGX H100 服务器为例，由于采用了 CPU+GPU 的异构架构，除了传统服务器中的主板、网卡板之外，**主要新增两种类型的高价质量 PCB 板**：承载 GPU 的 OAM（OCP Accelerator Module，加速卡模组）及实现 GPU 多卡互联的 UBB（Universal Baseboard，通用基板）。每颗 H100 芯片需要配备一张 OAM，共计 8 张 OAM 搭载至 1 张 UBB 之上，形成 8 卡互联。

图20：传统服务器 PCB 应用



资料来源：广合科技招股说明书，民生证券研究院

图21：DGX H100 AI 服务器中 OAM 和 UBB（绿色）

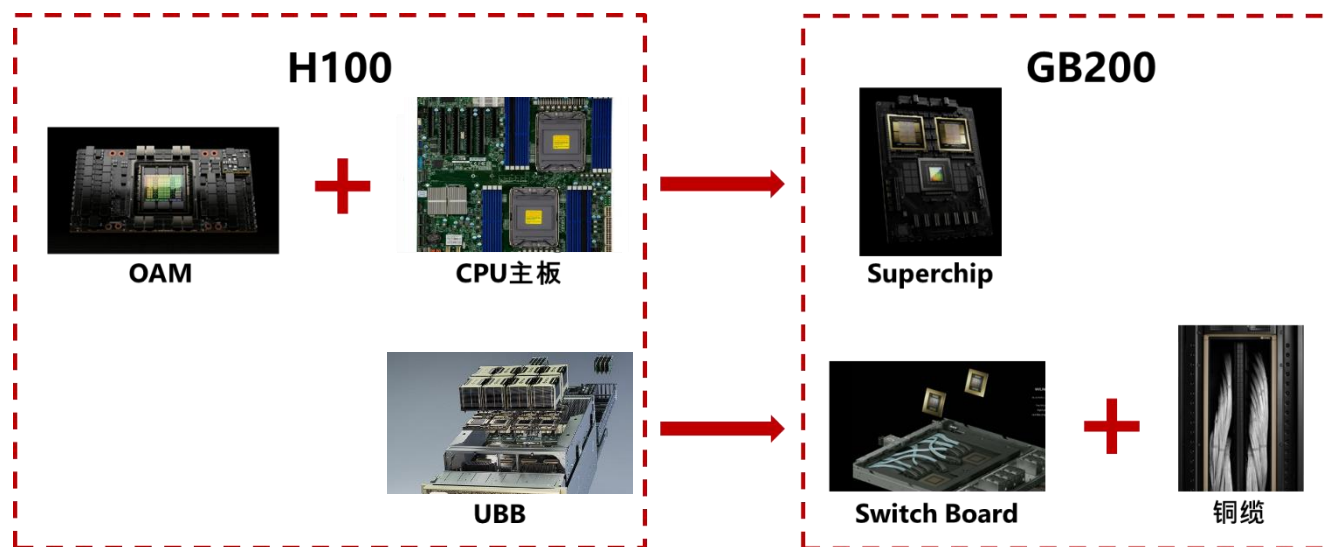


资料来源：英伟达，民生证券研究院

GB200 服务器整体架构与 DGX H100 相比发生显著变化。DGX H100 中每颗 GPU 芯片需要承载至一张 OAM 加速卡上，而 CPU 单独搭载至主板。而 GB200 中 1 颗 Grace CPU 及 2 颗 Blackwell GPU 直接搭载至 1 张 Superchip 板上，每个 Compute tray 中包含两张 Superchip 板，即 4 颗 GPU 及 2 颗 CPU。Superchip 板实现了 CPU 与 GPU 间的信号传输及芯片供电等功能，实际上替代了 DGX H100 中 OAM+CPU 主板的功能。**因此与 H100 相比，GB200 取消了主板，并且用一张大面积的 Superchip 板取代之前单颗 GPU 的 OAM。**

与 DGX H100 相比，GB200 由于取消了单台服务器内 8 颗 GPU 互联的设计，**因此不再需要 UBB 板**，一台 NVL36/72 机柜中，36/72 颗芯片的互联通过 Switch tray+铜缆实现，Switch tray 取代 UBB 承担了部分互联的功能。

图22: GB200 服务器 PCB 方案发生变化



资料来源: 英伟达, 太平洋科技, 民生证券研究院

当前谷歌、微软、亚马逊、Meta 四大海外云厂商中, 谷歌 TPU 及亚马逊 Trainium2 已形成大规模出货。与英伟达设计方案不同, 谷歌及亚马逊的 AI 服务器既没有采取类似 H100 的八卡架构, 也没有采取像类似 GB200 将 CPU 与 GPU 集成在一张 PCB 板的方案。如谷歌最新一代(第六代)的 TPU 芯片 Trillium, 将每四颗 TPU 搭载至一张 PCB 上, 组成 Compute Tray; CPU 单独搭载至另一张 PCB 中, 放置于另一个 Tray。载有 CPU 及 TPU 的 PCB 分别放置在不同的 Tray 中, 组合全新机架系统 TPU v6 Pod, 每个 TPU v6 Pod 由 512 个 Trillium 芯片组成, 提供高达 1.5 ExaFlops 的峰值性能,较上一代提升 83%。

图23: 搭载有谷歌第六代 TPU 芯片 Trillium 的 PCB

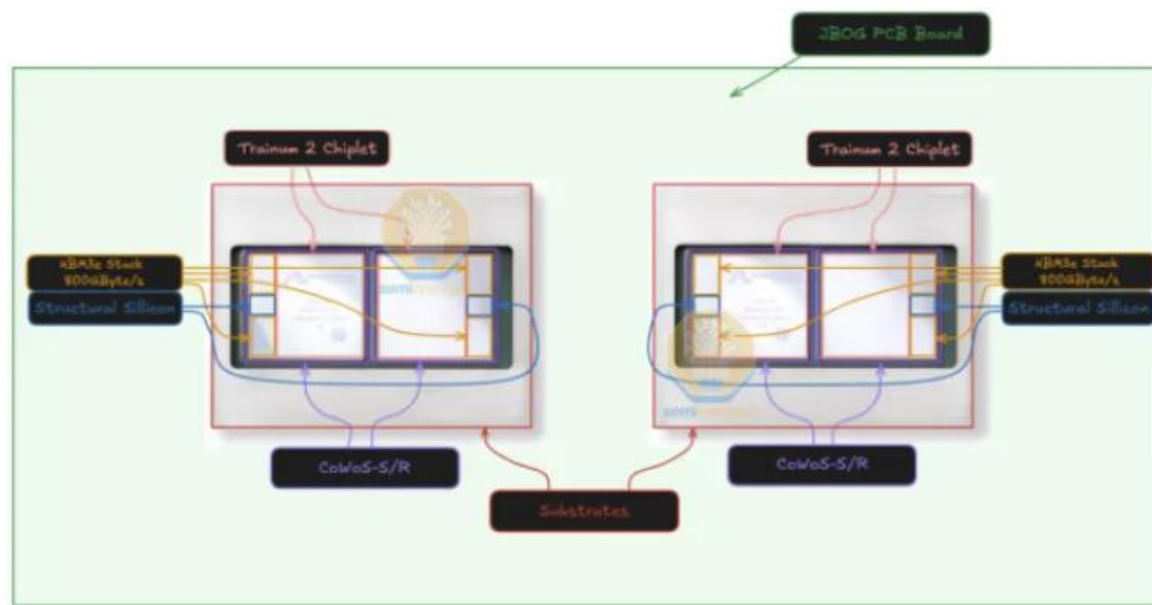


资料来源: Google , 半导体行业观察, 民生证券研究院

亚马逊同样采取了将 CPU 及 GPU 分离的设计方案, 两颗 Trainium2 芯片搭

载至一张 PCB 上, 组成 Compute Tray, 而两颗 CPU 单独搭载至另一张 PCB 上, 组成 Head Tray。此外, 柜内交换机为支持柜间信号互联的关键, 搭载交换机芯片的 PCB 亦具有较高价值量。

图24: 亚马逊 Trainium2 Compute Tray

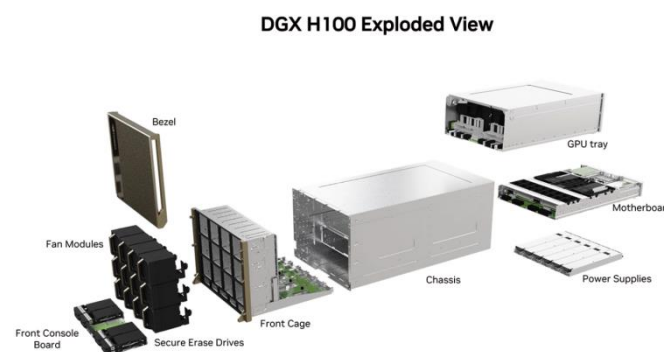


资料来源: Semianalysis, 民生证券研究院

2.2.2 铜缆配套, 机柜式服务器成为主流路径

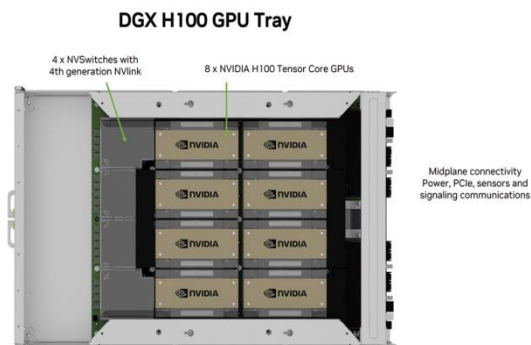
传统 AI 服务器主要是八卡架构, 谷歌等厂商的 TPU 服务器会采用四卡架构, 但是整体单个服务器内的 GPU 数量不会太多。以英伟达的 DGX H100 服务器为例, 服务器主要由 GPU Tray、Motherboard、电源等部件组成, 其中八张 GPU 内置在 GPU Tray 中, 通过 UBB 板上的 NV Switch 芯片互联, GPU 和 Switch 安装在同一个 Tray 中; Motherboard 上放置 CPU、内存、网卡等; Motherboard 和 GPU Tray 通过软板互联。GPU Tray 内部的八张 GPU 之间通过 NV Link 通信, 服务器和服务器之间通过光模块进行通信。

图25: 英伟达 DGX H100 服务器架构



资料来源: HPC Systems, 英伟达, 民生证券研究院

图26: 英伟达 DGX H100 GPU Tray



资料来源: HPC Systems, 英伟达, 民生证券研究院

英伟达推出的 GB200 NVL 72 机柜架构全面改变了传统的 AI 服务器形态。

在一台 NVL 72 中，机架最上层安放交换机，用于 Rack 之间的通信；机架次顶部和底部各安放 3 个（合计 6 个）33KW 的 Power Shelf，用于交流电到直流电的转换；机架中间部分是 18 个 Compute Tray 和 9 个 Switch Tray，分别用户安放 GB200 加速卡和 NV Switch 芯片，如果采用 NVL 36 的方案，则 Compute Tray 的数量将减少到 9 个。

图27：英伟达 GB200 NVL 72 机柜架构

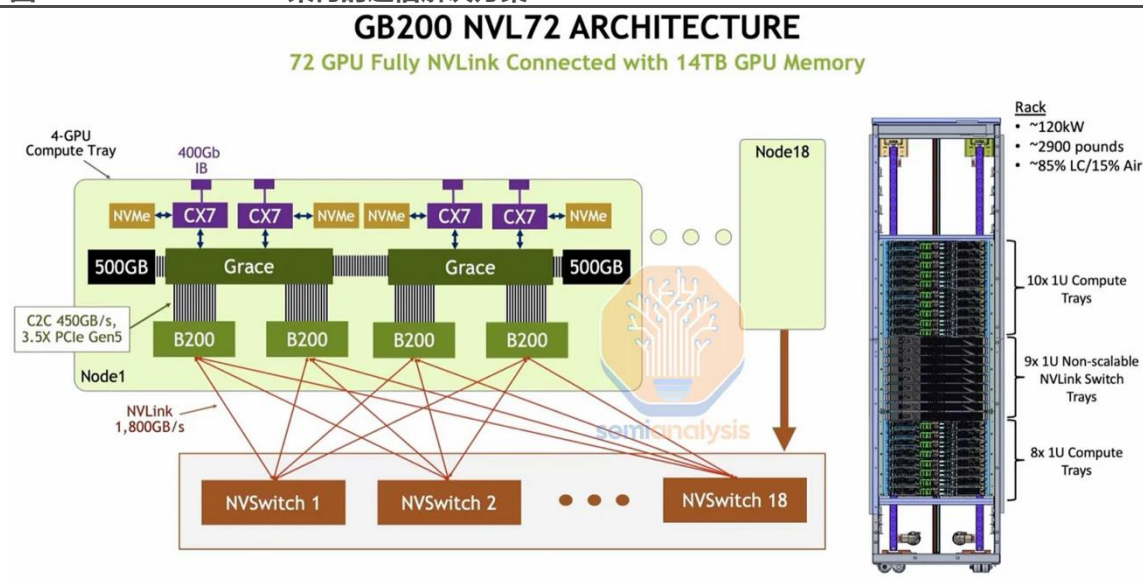


资料来源：华硕官网，民生证券研究院

高速铜缆是实现 GB200 机柜式服务器 36 卡或 72 卡之间通信的主要手段。

在 GB200 NVL72 的 Rack 内部，GPU Tray 到 Switch Tray 的互联采用高速铜缆，主要的优势在于，在传输距离较短的情况下，高速铜缆可以用更低成本获得更高的互联带宽。而其他的优势还包括：1) 功耗低：采用铜互联方案可以节省光电转换产生的能量损耗，单只 1.6T 光模块的功耗在 20W 左右，同时也降低了散热问题；2) 故障率低：光模块每年有 2%-5% 的损坏率，而铜连接更加稳定。

图28: GB200 NVL72 架构的通信解决方案



资料来源: Semianalysis, 民生证券研究院

铜缆在 GB200 机柜中的主要使用场景包含 Rack 内部以及 Rack 之间两个部分。其中 Rack 内部的铜互联主要负责 36 或 72 张加速卡之间的通信，最长的信号传输距离一般在 1 米以内。在 8 台机架之间的通信，英伟达提供了两种解决方案：1) 传统的光模块解决方案，在目前主流的方案下，通信速率一般为 1.8TB/s 的 NV Link 带宽的 1/4.5；2) 铜互联解决方案，可以以 1.8TB/s 的互联带宽进行通信，即将 576 张加速卡用铜互联的方案进行连接。

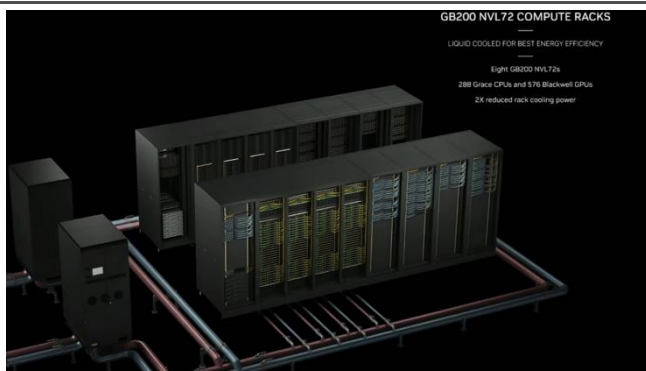
目前机柜间采用铜互联方案的渗透率仍较低，但考虑到机柜式 AI 服务器在性能和成本端的优势，预计未来主流的 AI 算力厂商也将推出机柜式服务器，高速铜缆在 AI 服务器内的渗透率有望逐步提升。

图29: Rack 内部铜互联解决方案



资料来源: 英伟达官网, 民生证券研究院

图30: Rack 之间的铜互联解决方案



资料来源: 英伟达官网, 民生证券研究院

2.2.3 AEC 解决下一代 AI 互联瓶颈

高速铜缆方案主要包括 DAC (无源铜缆)、ACC (有源铜缆)、AEC (有源

电缆)。过去机柜中主要采用 DAC 方案进行 AI 芯片间的互联，但目前 AI 芯片的互联带宽升级速度较快（以英伟达为例，公司每两年推出一代产品，每一代产品的互联带宽升级一倍），仅仅依靠 DAC 本身的升级无法满足 AI 芯片快速提升的互联带宽要求，通过 AEC 进行信号的重新定时和补偿，可以有效降低干扰，提升传输距离，减小线径。据 Lightcounting，2023 年至 2028 年 DAC 和 AEC 的市场空间 CAGR 增速分别为 25%和 45%，AEC 的市场空间增速显著更高。

1) DAC: 其由镀银铜导线和泡沫绝源芯片制成的高速电缆组成，因其不包含光电转化模块而具有较低的成本；

2) ACC: 其作为有源铜线，本质上是作为一根有源电缆来放大模拟信号。其利用 Redriver 芯片架构，并采用 CTLE 均衡来调整 Rx 端的增益；

3) AEC: 其代表了有源铜线电缆的一种更具创新性的方法。它在两头引入了具有时钟数据恢复功能的 retimer 芯片，对电信号重新定时和驱动，补偿铜缆的损耗，一般用于 7 米以内的传输距离；

4) AOC: 其由两端的两个模块组成，中间由一段光纤相连。光学模块和光缆都是集成的，两端的光学模块都需要激光组件。

图31：全球 DAC、AEC、AOC 市场规模（十亿美元）



资料来源：Lightcounting，民生证券研究院

具体而言，在短距离（<3m）、相对较低速率的传输场景下，DAC 的优势明显。主要原因是 DAC 两端无需增加 redriver 和 retimer 芯片，成本优势明显。然而随着速度和带宽的增加，DAC 的传输距离受到了一定的限制，有源电缆 AEC 在性能上的优势逐渐显现出来。在从 400G 到 800G 的过渡过程中，铜缆的损耗增大且互连长度无法满足需求，DAC 传输距离从 3 米缩短到 2 米。而 AEC 在保持低功耗和可负担性的同时得以满足中短距离传输需求，为短链路提供了一种经济高效的方式。

表3: DAC、AEC、AOC 性能对比

	DAC	AEC	AOC
400G 传输距离	< 3m	< 7m	< 300m
800G 传输距离	< 2m	< 2.5m	< 300m
功耗	低	低	高
费用	低	中等	高
传输速度	快	快	慢

资料来源: Asterfusion, 民生证券研究院

DAC、ACC、AEC 和 AOC 在 AI 服务器中均有采用, 而 AEC 有望伴随云厂商自研 AI 芯片的市占率提升加速渗透。目前市场主流的机柜式 AI 服务器中, 英伟达的 GB200 根据不同的距离和带宽需求, 在机柜内部和机柜间采用 DAC 和 ACC 进行互联, 服务器和交换机之间则主要采用 AOC 进行互联。而在微软和亚马逊的自研 AI 服务器中, AEC 则替代了 DAC, 成为机柜内互联 AI 芯片的主要产品。不同厂商采用不同解决方案的主要衡量因素是性能、成本、互联带宽等。

图32: DAC、AEC、AOC 功耗对比

MAX POWER (W)

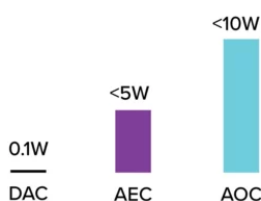


Figure 1 Comparaison de la consommation électrique du câble à 400G

资料来源: Precision OT, 民生证券研究院

图33: DAC、AEC、AOC 成本对比

COST (\$)

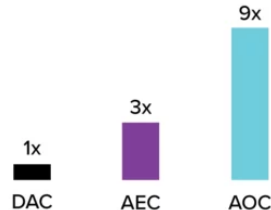


Figure 3 Approximatif Comparaison des coûts DAC/AEC/AOC

资料来源: Precision OT, 民生证券研究院

2.2.4 AEC 渗透率有望提升, 以 AWS 服务器为例

类似英伟达使用 NVLink 扩展网络实现芯片互联, Trainium2 采用 NeuronLink 超速高带宽、低延迟的芯片间互联技术实现芯片与芯片之间的高速互联, 支持更多的芯片连接和更高的数据传输速率, 为大规模集群训练和推理提供了强大的扩展性。Trainium2 提供两种产品规格, 分别采用了 2D Torus 和 3D Torus 的网络拓扑结构, 可将 16 个 (4×4) 或 64 个 (4×4×4) Trainium2 芯片连接在一起, 形成不同规模的计算集群, 以适应不同的计算需求。在两种服务器架构中, PCB 连接同一个 Compute Tray 内部的 2 颗加速卡, 其他加速卡之间的互联, 以及机柜和机柜间的互联则采用 AEC。

1) 16 卡服务器: AWS 基于 Trainium2 打造了 AI 服务器, 每个 Trainium2 服务器占用 18 个机架单元 (18U), 由 1 个 2U 的 CPU Head Tray 和 8 个与之相连的 2U Trainium2 Compute Tray 组成。每个 Compute Tray 搭载两个 Trainium2 芯片, 与英伟达 GB200 的架构不同, 其 Compute Tray 中未搭载 CPU,

需要依托 Head Tray 运行；两个 CPU 单独搭载至一个 Head Tray 中，通过外部 PCIe 5.0 x16 DAC 无源铜缆与 8 个 Compute Tray 相连。每台 Trainium2 服务器整合了 16 颗 Trainium2 芯片，具备 20.8 PFLOPS 的算力，适合数十亿参数大语言模型的训练与部署；

2) 64 卡服务器：AWS 还基于 Trainium2 进一步打造了 Trainium2 Ultra 机架方案，其集成 64 颗 Trainium2 芯片，将 4 台 Trn2 服务器连接成 1 台巨型服务器，可提供相比当前 EC2 AI 服务器多达 5 倍的算力和 10 倍的内存，FP8 峰值算力可达 83.2PFLOPS，能够支撑万亿参数 AI 模型的实时推理性能。

图34: Trainium2 Rack



资料来源: SemiAnalysis, 民生证券研究院

图35: Trainium2-Ultra Rack



资料来源: SemiAnalysis, 民生证券研究院

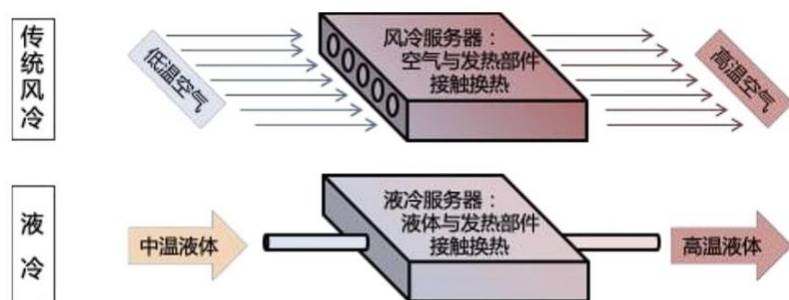
2.2.5 散热：芯片功率提高，液冷时代已至

传统风冷是以空气为热量传输媒介，液冷技术是将高比热容的液体作为热量传输媒介，直接或间接接触发热器件，缩短送风距离，传热路径短，换热效率高。

AI 算力芯片计算能力强、功耗大，产生热量较多，若不能及时进行散热以实现芯片的降温，轻则影响芯片性能，重则将使芯片使用寿命减少乃至损伤、报废。传统数据中心中，由于以 CPU 云计算为主，芯片功耗较低，因此一般采取风冷散热。然而随着 AI 算力芯片功耗密度大幅提高，传统风冷散热技术已难以满足当前的高密度计算散热需求，诸如英伟达 GB200 NVL72、谷歌 TPU 等需要采取散热效果更好的液冷辅助散热，数据中心液冷时代已至。

根据 IDC 数据，中国液冷服务器市场 2024 上半年同比大幅增长 98.3%，市场规模达到 12.6 亿美元，出货量同比增长 81.8%，**预计 2023-2028 年，中国液冷服务器年复合增长率将达 47.6%**，市场规模有望在 2028 年达到 102 亿美元。

图36：风冷及液冷对比



资料来源：赛迪顾问《中国液冷数据中心发展白皮书》，民生证券研究院

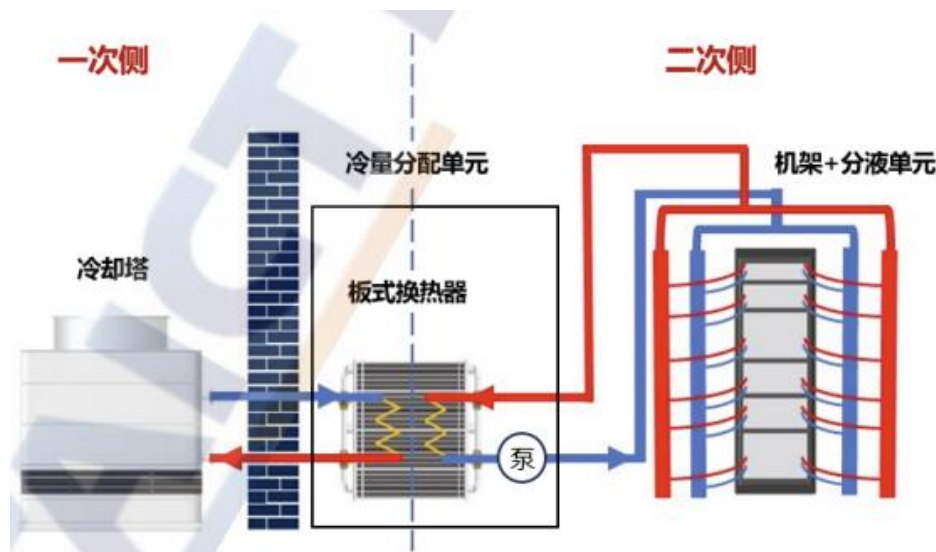
液冷技术主要可分为非接触式液冷及接触式液冷两类。其中非接触式液冷主要指冷板式液冷，将服务器发热元件（CPU/GPU/DIMM 等）贴近冷板，通过液冷板将发热器件的热量传递给封闭在循环管路中的冷却液体，以带走热量。

冷板式液冷对于服务器整体改动较小，主要可通过加装液冷模块、采用集中式或分布式 CDU 供液、Manifold 分液，对芯片、内存等部件进行精准制冷；相较其他液冷方案，冷板式液冷在可靠性、可维护性、技术成熟度、适用性等方面具有优势，利于算力中心机房改造，根据 IDC，冷板式液冷成为当前绝对主导的液冷解决方案，占据国内市场 95%以上份额。

冷板式液冷系统可分为一次侧（室外）循环和二次侧循环（室内）两部分。一次侧系统主要由室外散热单元、一次侧水泵、定压补水装置和管路等部件构成。二次侧系统主要由换热冷板、热交换单元和循环管路、冷源等部件构成。一次侧的热量转移主要是通过水温的升降实现，二次侧循环主要通过冷却液温度的升降实现热量转移。

换热冷板及 CDU 亦为液冷散热系统的重要组成部分。换热冷板常作为电子设备的底座或顶板，通过空气、水或其他冷却介质在通道中的强迫对流，带走服务器中的耗散热。冷量分配单元(Coolant Distribution Unit, CDU)可以看作室内机与室外机的连接点，由板式换热器、电动比例阀、二次侧循环泵膨胀罐、安全阀、进出水管专用接头、控制器及其面板等部件组成

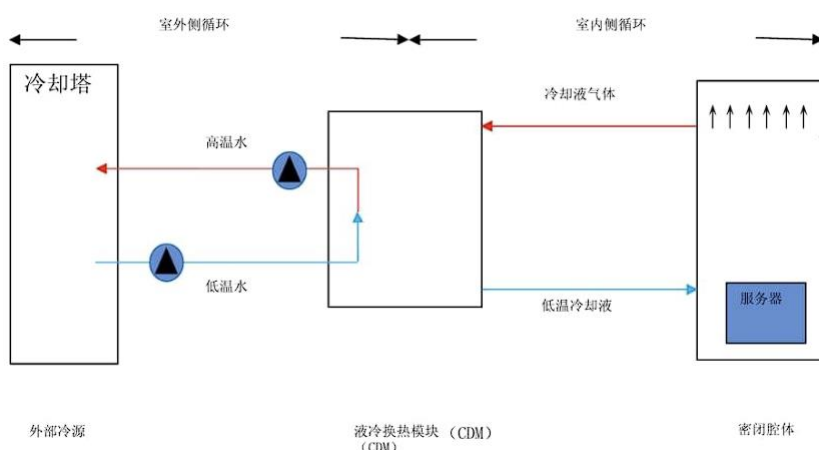
图37：冷板式液冷系统示意图



资料来源：中国信通院，民生证券研究院

接触式液冷的液体与发热源直接接触，包括浸没式液冷和喷淋式液冷两种。浸没式液冷是将服务器浸没在冷却液中，通过液体温升或相变带走服务器中所有发热元件的热量。喷淋式液冷的冷却液从服务器机箱顶部的喷淋模块滴下来，通过冷却液与发热元件之间的接触进行对流换热，从而为发热元件降温，再通过服务器内的流道汇集至换热器将热量散发。相比非接触式液冷，接触式液冷可完全去除散热风扇，散热及节能效果更好，数据中心 PUE 值可降至 1.1 及以下，但相应对服务器机柜及机房配套设施的投入及改造成本更高，运维难度更大。未来随着浸没式液冷技术标准化推进、应用部署成本降低，有望加速大规模商用进展。

图38：浸没式液冷原理示意图

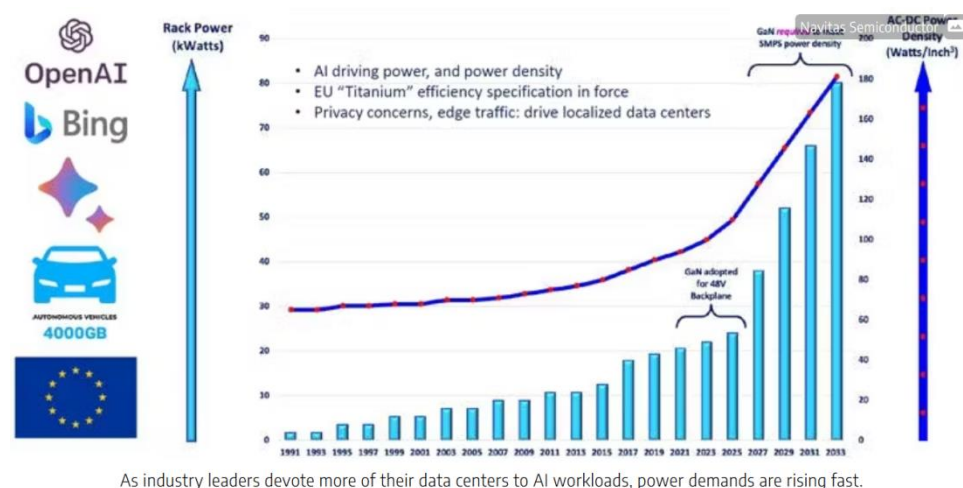


资料来源：赛迪顾问《中国液冷数据中心发展白皮书》，民生证券研究院

2.2.6 电源：功率密度不断提升，电源要求持续升级

英伟达率先推出机柜式服务器，其他 CSP 厂商纷纷跟进，AI 服务器功率密度快速提升。AI 服务器的功率密度提升来自两个方面：1) 更高算力的加速卡需要更高的功耗：以英伟达的产品为例，2020 年推出的 A100 加速卡单卡功耗为 300W，而 2024 年推出的 GB200 加速卡单卡功耗则达到了 1000W；2) 机柜内集成的加速卡数量不断提升：传统 AI 服务器以 8 卡形式为主，而进入 2024 年，英伟达推出了机柜架构的 AI 服务器，大大提升了单台服务器在算力、通信等方面的性能，其他厂商也纷纷开始拥抱机柜式 AI 服务器的浪潮，目前主流 CSP 厂商均已推出或在研机柜式 AI 服务器。展望未来，伴随算力供应商产品的不断迭代，单台 AI 服务器内集成的加速卡数量有望持续提升。

图39：AI 带动服务器机柜功率及功率密度持续提升

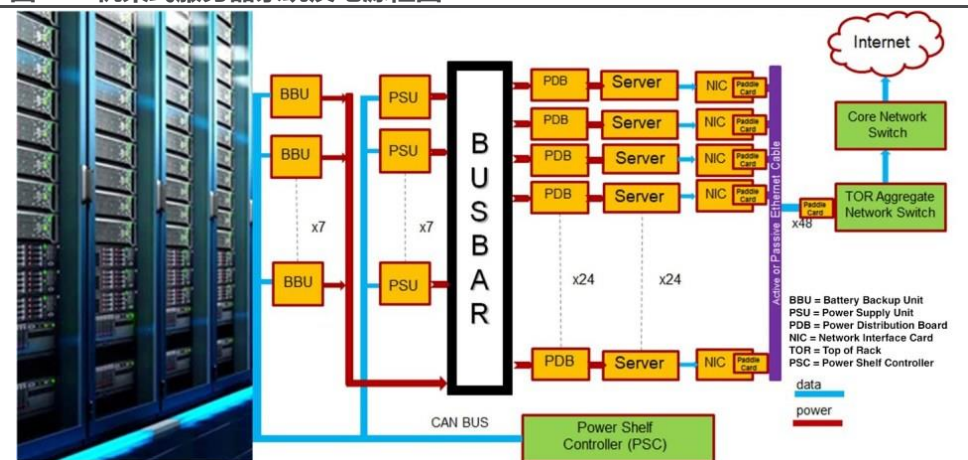


As industry leaders devote more of their data centers to AI workloads, power demands are rising fast.

资料来源：Electronic Design，民生证券研究院

服务器系统的可靠性要求较高，因此电源设计时需要考虑冗余。在 AI 服务器工作过程中，如果电源出现故障，则可能造成系统宕机，数据丢失等问题，为了避免损失，AI 服务器设计中需要考虑冗余。AI 服务器系统通常采用 N+1 或 N+N 冗余设计，意味着服务器电源功率要大于系统中所有耗电元器件，如 GPU 和 CPU 的功率总和，对电源提出更多要求。在选择 AI 服务器电源冗余时，通常考虑系统成本以及可靠性要求进行选择。

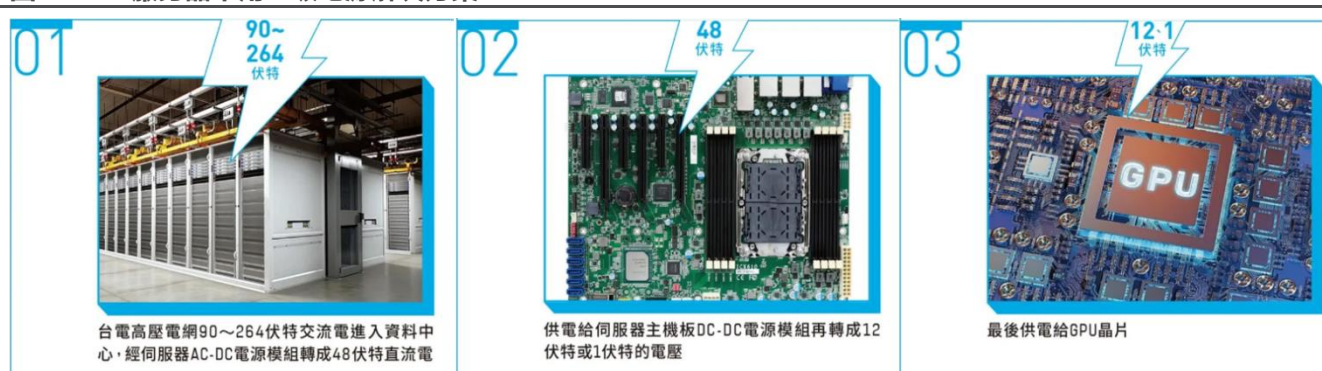
图40：机架式服务器系统及电源框图



资料来源：德州仪器（TI），民生证券研究院

AI 服务器的升级对电源供电要求提出了更多挑战，目前机柜式服务器主流采用三级电源供电。例如英伟达的 GB200 机柜式 AI 服务器，其中一级电源为 AC-DC（交流转直流），使用 power shelf 将电网中的交流电转换为机柜运行所使用的直流电；二级电源为 DC-DC（高压直流转低压直流），以 PDB（Power Distribution Board）的形式内置于 compute tray，将 48V 直流电降压为 12V 直流电；三级电源也是 DC-DC（将低压直流进一步降压），体积较小，位于 Superchip 内部（芯片周围），将二次电源转化后的 12V 直流电进一步降压为 0.8V 左右直流电，用于 GPU 和 CPU 芯片供电。

图41：AI 服务器采用三级电源解决方案



资料来源：台达，雅虎，民生证券研究院

3 端侧：豆包出圈，互联网巨头入局 AI 终端

3.1 字节豆包先行，加速端侧落地

3.1.1 云厂商合作伙伴+终端落地将是 25 年的主要趋势

当前大模型厂商主要分为互联网巨头和第三方科技公司两大阵营：

(1) 互联网巨头：代表如字节、阿里、百度、腾讯等，其优势在于广大的用户群体和流量变现能力。这些互联网巨头旗下均有高活跃度应用可以帮助引流，如字节的抖音、阿里的淘宝、百度的搜索引擎、腾讯的微信等，同时，这些应用还积累了大量可用于模型训练的私域数据。

(2) 第三方科技公司：代表如智谱、Kimi、阶跃星辰，主要依托模型的特色功能出圈，如 Kimi 在长上下文窗口技术上取得突破，率先支持 200 万字上下文长度。尽管这些独角兽具有较强的技术实力，但由于缺少可以承载模型能力的下游应用场景，因此只能寻求和第三方软硬件公司的合作。

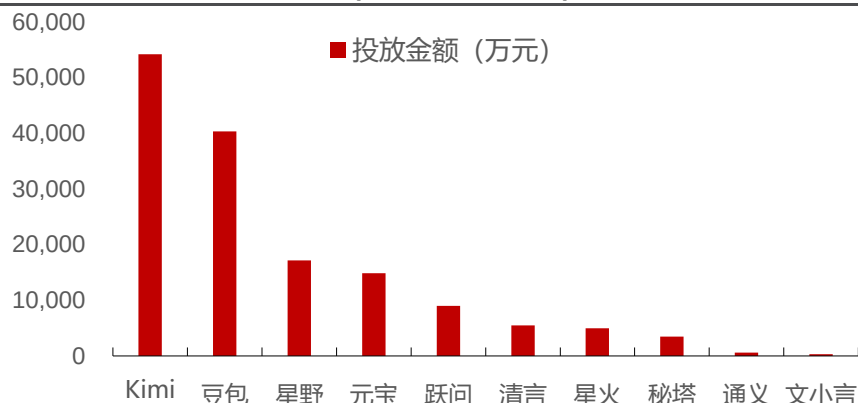
我们认为 25 年有望迎来云端和终端的共振，而二者的交集正是字节产业链，建议关注字节火山引擎合作伙伴以及 AI 终端落地节奏及有望受益于此的供应链合作伙伴。

3.1.2 大模型千帆竞渡，字节豆包为何脱颖而出

在互联网巨头阵营中，字节目前遥遥领先于同业。相较于竞争对手而言，字节的 AI 起步相对较晚，但后来居上，国内百度、阿里、商汤、科大讯飞在去年 3 月开始都陆续推出了大模型新品和 AI 应用，而字节的豆包则是在同年 8 月才发布。尽管如此，今年 5-7 月，豆包 App 日新增用户从 20 万迅速飙升至 90 万，并在 9 月率先成为国内用户规模破亿的首个 AI 应用。据量子位智库数据，截至 11 月底，豆包 2024 年的累计用户规模已超过 1.6 亿；11 月平均每天有 80 万新用户下载豆包，单日活跃用户近 900 万，位居 AI 应用全球第二、国内第一。

我们认为算力资源充足和愿意大力投入是字节豆包迅速起量的主要原因。算力资源方面，字节旗下火山引擎支持多芯、多云架构，拥有超大规模算力，支持万卡集群组网、万亿参数 MoE 大模型；提供超高性能网络，支持 3.2Tbps RDMA 网络，全球网络 POP 覆盖广，时延优化最高达 75%；目前花钱投流买量已成为 AI 产品启动的最快捷的方式之一，据 AppGrowing，截至 11 月 15 日，Kimi、豆包、星野等国内十款大模型产品，今年合计已投放超 625 万条广告，投放金额达 15 亿元，其中，豆包以 4 亿+元位列第二。在各家的投放渠道中，基本都离不开字节的巨量引擎（字节旗下广告投放平台，涵盖今日头条、抖音、西瓜视频等营销资源），而背靠字节的豆包，更是将流量池的优势发挥到了极致。

图42：国内各大模型 2024 年（截至 11 月 15 日）广告投放投入



资料来源：AppGrowing，民生证券研究院

除广告投入外，字节还积极布局 AI 硬件，为豆包寻找 C 端应用场景，打开市场空间。2024 年 9 月，字节完成了对 OWS 开放式耳机品牌 Oladance 的全资收购，进一步拓展其硬件产品，以完善 AI 服务生态。10 月，其收购后的首款产品 Ola Friend 智能体耳机正式发布，该产品功能上主要特征是接入了豆包 AI 大模型，为用户提供陪伴式体验。此外，多款第三方硬件产品也接入豆包大模型，目前 OPPO、vivo、荣耀、小米、三星、华硕已联合火山引擎发起智能终端大模型联盟，OPPO 小布助手、荣耀 MagicBook 的 YOYO 助理、小米的小爱同学，以及华硕笔记本电脑的豆叮 AI 助手等应用，均已接入豆包大模型服务。

图43：字节跳动 2024 年硬件布局



资料来源：Tech 星球公众号，腾讯新闻等，民生证券研究院整理

3.1.3 字节豆包算力需求测算

据 AI 产品榜，字节豆包目前已经成为除了 OpenAI ChatGPT 外月活数量最高的 AI 大模型。截止 2024 年末，豆包大模型 MAU 达到 5998 万人，后续有望持续提升，我们假设在保守、中性、乐观三种情况下，2025-2026 年豆包大模型

MAU 分别达到 1 亿、1.5 亿、2 亿人次，云雀大模型参数量仍为 1300 亿，则豆包大模型等效 H20 算力需求将分别达到 72、108、181 万张，对应 AI 服务器需求分别达到 759、1139、1898 亿元。

表4：豆包大模型推理需求测算

豆包大模型推理算力需求测算	2024 年末	2025-2026E		
		保守	中性	乐观
MAU (万人)	6000	10000	15000	20000
DAU (万人)	900	1500	2250	3000
豆包 (云雀) 大模型参数量	1300	1300	1300	1300
日均 token 调用量 (亿)	40000	66667	100000	133333
推理计算时间 (s)	2	2	2	2
日均每秒 token 计算数量 (亿)	0.31	0.51	0.77	1.03
峰值倍数	4	4	4	5
算力需求公式	云端 AI 推理算力需求 $\approx 2 \times$ 模型参数量 $\times$ 数据规模 $\times$ 峰值倍数			
算力需求结果 (FLOPS)	3.20988E+19	5.34979E+19	8.02469E+19	1.33745E+20
H20 单卡算力 (TFLOPS)	148	148	148	148
MFU	50%	50%	50%	50%
需要 H20 卡数量 (万张)	43	72	108	181
H20 单价 (万元)	8.4	8.4	8.4	8.4
H20 服务器均价 (万元)	84	84	84	84
豆包大模型创造 AI 服务器需求 (亿元)	455.46	759.09	1138.64	1897.73

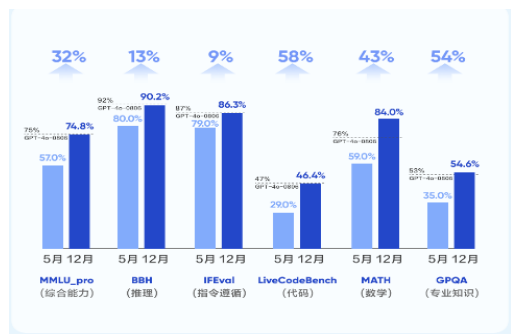
资料来源：豆包官网，OpenAI 官网，英伟达官网，36 氪，民生证券研究院整理

3.1.4 火山引擎大会来袭，有望成为字节硬件布局宣传窗口

字节目前积极合作供应链企业，领导行业探索未来 AI 落地新方向，火山引擎大会有望成为硬件布局的宣传窗口。此前，在 5 月的火山引擎春季大会上，字节对外展示 3 款 AI 硬件的合作产品，包括蔚蓝机器狗、听力熊学习机和一款学习机器人。现在，火山引擎将于 12 月 18-19 日举办冬季《FORCE 原动力大会》。大会将展示豆包大模型家族的全新升级，拓展 AI 应用场景边界，并打造 2000 平方米 AI 展区，互动演绎 AI 未来：

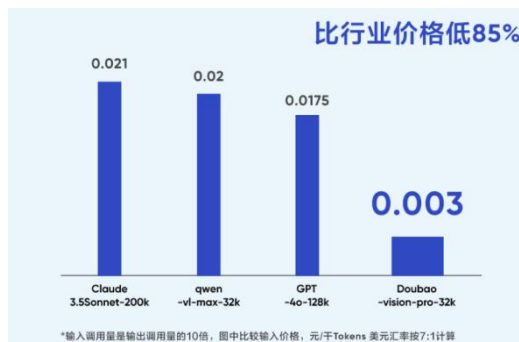
- 12 月 18 日，字节跳动正式发布了豆包视觉理解模型，该模型具备更强的内容识别能力、理解和推理、视觉描述等能力。该模型输入价格为 0.003 元/千 tokens，比行业价格低 85%，视觉理解模型进入“厘时代”。此外，大会将展现火山引擎的全栈 AI 能力，还将进一步结合边缘域、汽车、消费等领域实体产业，诠释如何实现 AI+。乐鑫科技也将参与此次大会，并发表主题演讲。此外，豆包通用模型 pro 完成新版本迭代，综合任务处理能力较 5 月提升 32%，在推理上提升 13%，在指令遵循上提升 9%，在代码上提升 58%，在数学上提升 43%等。

图44：豆包通用模型 pro 能力全面提升



资料来源：火山引擎公众号，民生证券研究院

图45：Trainium2-Ultra Rack



资料来源：火山引擎公众号，民生证券研究院

- 12月19日，大会的开发者论坛亮相了火山方舟、扣子、豆包 MarsCode 等产品。其中，火山方舟发布了高代码智能体和 API 接口，助力开发者高效调用大模型；扣子发布 1.5 版本，提供了全新的应用开发环境，全面升级多模态能力，为开发者提供专业模板，目前已拥有超过 100 万的活跃开发者，成为新一代精品应用开发平台。

图46：火山引擎时间安排表



资料来源：火山引擎，民生证券研究院

豆包引发市场关注热潮，与其相关的产业链上下游企业同样受到高度关注，国产算力+端侧产业链全面崛起，为字节等厂商提供更有力的供应链支撑，保驾护航。

表5：国产算力+端侧产业链公司

领域	公司	业务
AI算力	润泽科技	字节数据中心供应商。
AI端侧	恒玄科技	提供SoC芯片， BES2700芯片进入字节ola耳机。
	乐鑫科技	提供SoC芯片， esp32芯片被用于AI玩具“显眼包”。
	中科蓝讯	提供SoC芯片， 公司BT895x芯片目前已经和火山方舟MaaS平台对接。
	歌尔股份	提供代工服务， 字节跳动子公司PICO与歌尔股份签署长期合作协议。
	炬芯科技	提供SoC芯片， ATS3085L芯片已用于荣耀手环9。

全志科技	提供SoC芯片，MR527芯片用于石头V20扫地机器人。
星辰科技	提供SoC芯片，计划2025年推出AI眼镜芯片。
瑞芯微	提供SoC芯片，公司边缘AI芯片广泛应用于智能手机、智能家居、汽车智能座舱等多个领域。
国光电器	提供一站式电声解决方案，将VR/AR作为业务发展重点，已布局AR眼镜等领域。
中科创达	智能操作系统及端侧智能技术和产品提供商，与火山引擎共建人工智能大模型联合实验室。
云天励飞	具备端到端整体解决方案的AI公司，自研了DeepEdge系列边缘人工智能芯片满足部署的需求。

资料来源：乐鑫董办公众号，每日经济新闻等，民生证券研究院整理

以豆包为代表的 AI 产品，正以前所未有的速度在教育、医疗、文娱等领域广泛应用，不仅深刻改变着行业的传统格局，更成为推动包括 AI 终端在内的各领域实现变革与进步的强大引擎。

3.2 AI 终端空间广阔，SoC 是影响体验的核心硬件

3.2.1 AI 终端不断扩容，市场空间广阔

AI 赋能有望改变电子产业的增长曲线，未来广阔的终端硬件都有重估潜力。

22 年全球智能音箱市场出货量为 1.2 亿台；23 年全球品牌 TWS 耳机年销量约 3 亿对；2023 年手表/手环年销量预计达 1.61 亿只；23 年中国戴眼镜人群接近 7 亿，在引入 AI 功能后，有望带动传统产品升级，刺激换机需求。此外，据 Contrive Datum Insights 数据，2030 年全球 AI 玩具市场规模有望达 351.1 亿美元。

图47：AI 终端广阔空间

主要场景	智能可穿戴		智能办公设备		智能家居	
示意图						
主要IOT产品	TWS耳机	智能手表/手环	平板电脑	PC	TV	智能音箱
2023年出货量 (亿)	2.95	1.61	1.33	2.5	1.96	1.2 (2022年)
TOP1品牌出货量	苹果: 0.82	苹果: 0.4	苹果: 0.52	联想: 0.59	三星: 0.36	亚马逊: 0.34
TOP2	三星: 0.25	华为: 0.13	三星: 0.26	惠普: 0.53	海信: 0.26	谷歌: 0.20
TOP3	boAt: 0.17	三星: 0.12	联想: 0.09	戴尔: 0.40	TCL: 0.26	苹果: 0.15
TOP4	小米: 0.16	Noise: 0.09	华为: 0.08	苹果: 0.23	LGE: 0.21	阿里巴巴: 0.15
创新方向	AI耳机、OWS耳机；带摄像头的耳机等	健康监测功能和电池续航	折叠屏、手写笔、外接键盘等	AI PC	智能电视、智能投影等	语音、屏幕多场景智能交互中心

资料来源：Canalys，IDC，AVC Revo，Counterpoint Research 等，民生证券研究院整理

我们看好 AI+智能终端的趋势，AI 将重构电子产业的成长，为智能硬件注入全新的活力，带来产品逻辑的深度变革，加速硬件的智能化、伴侣化趋势。

无论是手机、PC、AIOT、可穿戴设备、汽车电子，都有重估的潜力。当下，各大厂商纷纷布局，应用端革新渐渐开：

1) **手机端**：当前系统级 AI+打通第三方 App 确立为端侧 AI 的发展方向，AI

终端成为长期产业趋势。苹果方面，公司于 10 月 29 日正式推出 iOS 18.1 并面向北美用户推出 Apple Intelligence；安卓方面，智谱于 10 月 25 日发布的 AI 智能体 AutoGLM 也开启内测；

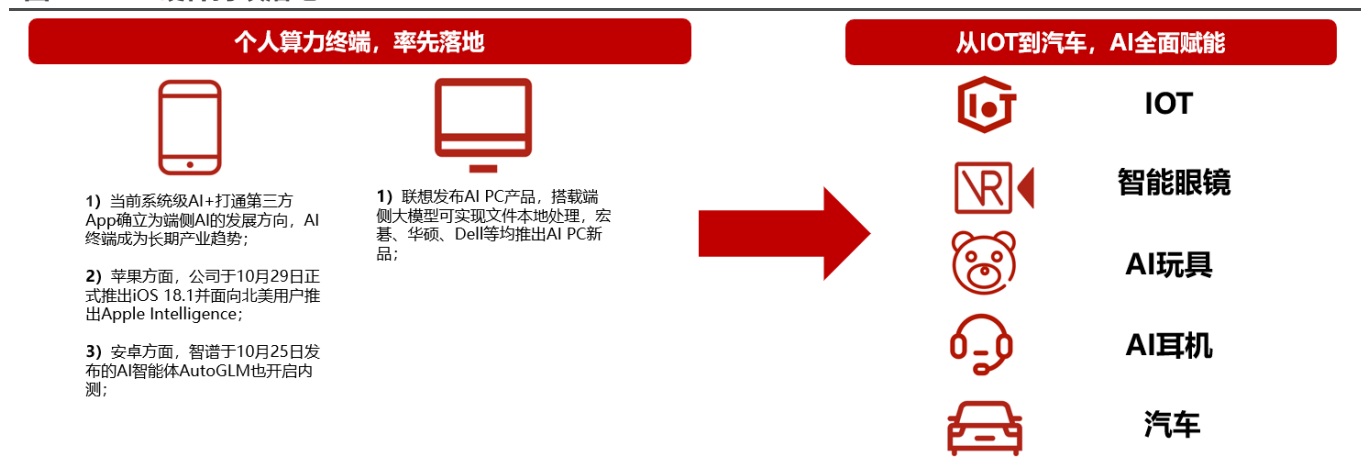
2) PC 端：联想发布 AI PC 产品，搭载端侧大模型可实现文件本地处理，宏碁、华硕、Dell 等均推出 AI PC 新品；

3) AIOT：智能音箱方面，阿里巴巴天猫精灵，接入“通义千问”大模型，提升智能交互体验；AI 玩具方面，字节基于大模型开发情感陪伴玩偶“显眼包”，具备语音交互、语音识别等功能，满足用户情感陪伴。

4) 可穿戴设备：AI 眼镜方面，雷朋联合 Meta 推出 Ray-Ban 智能眼镜，搭载 Llama 模型，实现拍摄、识别及翻译外部物体、并可与用户进行语音交互；AI 耳机方面，字节推出 Ola Friend 智能体耳机，产品接入豆包 AI 大模型，为用户提供陪伴式体验；

5) 智能汽车：小米汽车的端到端大模型全场景智能驾驶，预计将于 24 年 12 月底正式发布；特斯拉的 FSD 计划 25 年 Q1 在中国和欧洲推出。

图48：AI+硬件持续落地



资料来源：科创板日报，京报网等，民生证券研究院整理

在广阔的潜在市场需求下，我们看好字节等头部大模型公司与各领域硬件厂商持续加深合作，持续推出智能对话、儿童早教等多样化创新应用，新型智能终端即将迎来规模落地。

3.2.2 数字芯片厂商有望高度受益于 AI 终端浪潮

核心云/硬件厂商的重要“合作伙伴”将成为投资主线。随着 AI 产业开始走向应用层/推理侧，AI 产业的话语权或将向云厂商/电子品牌商倾斜，但不论话语权如何交接，AI 应用的落地都需要硬件的承载，在此过程中成为行业领先者的重要“合作伙伴”至关重要，重要云厂商（谷歌、微软、Meta、字节、百度）+电子

品牌厂商（苹果、华为、小米、特斯拉等）合作伙伴值得关注。

表6：数字芯片公司合作伙伴梳理

公司	下游核心客户
全志科技	天猫精灵、小米、小度等品牌智能音箱； 石头、云鲸、小米、追觅等品牌扫地机； 吉利、红旗、五菱等汽车品牌
晶晨股份	小米、阿里巴巴、Google、Amazon、创维、中兴通讯等； SONOS、三星、JBL； 宝马、林肯、Jeep、沃尔沃、极氪等
瑞芯微	阿里、小米、百度、安克创新、腾讯、网易、科沃斯等； 比亚迪、广汽、汇川等
中科蓝讯	小米、realme、百度、万魔、倍思、Anker、漫步者、传音、boAt、Noise、科大讯飞、TCL等
恒玄科技	三星、OPPO、小米、荣耀、华为、vivo 等； 哈曼、安克创新、漫步者、韶音等； 阿里、百度、字节跳动、谷歌等
炬芯科技	哈曼、SONY、安克创新、荣耀、小米、罗技、Razer、漫步者等
乐鑫科技	小米、谷歌、字节跳动等
富瀚微	专业安防客户：海康、大华等； AIoT 市场； 移动等三大运营商、萤石等
泰凌微	谷歌、亚马逊、小米、罗技、联想、JBL、索尼等

资料来源：各公司公告，民生证券研究院整理

在 AI 应用落地成为主叙事方向的情况下，零部件供应链企业将受益于终端产品的销量快速增长，将拥有较大的业绩弹性空间。其中，SoC 作为 AI 硬件的核心部分，重要性也愈加凸显。

万变不离其宗，SOC 能力不断提升。复盘近两年 AI 硬件的几轮投资热潮，从去年初的智能音箱到年中 AI PC、年末 MR，从今年上半年的 AI Phone 到下半年的 AI 耳机、AI 眼镜、AI 玩具，硬件形态快速创新，但其对核心 SOC 的能力要求仍主要围绕连接+处理两个方向进化，优秀 SOC 厂商的能力将不断被证明。

依据各个赛道硬件的特点，SoC 将具备不同的需求。1) 可穿戴赛道：连接+低功耗是核心痛点，集成 ISP 等多模态处理能力是重点方向；2) AI 潮玩等长尾市场：稳定连接+高性价比是核心需求；3) 端侧模型部署：多核异构（CPU+GPU+NPU）或 ASIC 芯片是主流方案。

以 AI 眼镜为例，我们认为前期雷朋眼镜的热销主要得益于其 3A 的独到之处，即自动曝光、自动对焦、自动白平衡，由于 AI 眼镜无法像手机一样通过人手实时操控调整对焦，所以对其智能化的物体识别等效果会更为重要，这也是多模态的基础。在此过程中，ISP 芯片的视觉处理能力发挥了至关重要的作用。在早期耳机时代，连接和音频处理是重中之重，但伴随着耳机向眼镜过渡，视觉处理则更为关键。

当下，SoC 行业走出底部+下游 AI 智能硬件需求提升，诸多厂商有望在 AI

浪潮下走出新的成长曲线：

- ✓ **恒玄科技**：2024H1 公司新一代智能可穿戴芯片 BES2800 实现量产出货，芯片基于 6nm FinFET 工艺集成了 Wi-Fi6，可实现超低功耗无线连接；
- ✓ **乐鑫科技**：ESP32-S3 已可对接 OpenAI 的 ChatGPT 或百度“文心一言”等云端 AI 应用，新产品 ESP32-P4 具备边缘 AI 功能；
- ✓ **中科蓝讯**：公司讯龙三代 BT896X 系列芯片已应用在百度新推出的小度添添 AI 平板机器人的智能音箱中，可实现 AI 语音交互。
- ✓ **晶晨股份**：2024H1 公司基于新一代 ARM V9 架构和自主研发边缘 AI 能力的 6nm 商用芯片流片成功，并已获得首批商用订单；
- ✓ **瑞芯微**：公司 RK3588、RK3576 采用高性能 CPU 和 GPU 内核并带有 6T NPU 处理单元，针对端侧主流的 2B 参数数量级别的模型运行速度能达到每秒生成 10 token 以上，满足小模型在边、端侧部署的需求；

表7：主要算力芯片参数

公司	产品型号	内核	主要算力芯片 主频	算力	制程
恒玄科技	BES2800	双核ARM Cortex-M55	300MHz	-	6nm
晶晨股份	A311D	4*Cortex-A73 + 2*Cortex-A53	最高2.2GHz	5 TOPS	12nm(6nm流片成功)
瑞芯微	RK3588	4*Cortex-A76 + 4*Cortex-A55	最高2.4GHz	6 TOPS	8nm
北京君正	T41	双核XBurst2	1.2~1.4GHz	1.2 TOPS	12nm
全志科技	V853	双核Cortex-A7 + RISC-V E907	1GHz+600MHz	1~2 TOPS	22nm
富瀚微	FH8898	4 核RISC 处理器	-	2 TOPS	22nm
中科蓝讯	BT8952F	RISC-V+DSP扩展	125MHz+270MHz	-	22nm
炬芯科技	ATS283XP	32bits RISC+DSP扩展	264MHz	-	40nm
乐鑫科技	ESP32-S3	Xtensa 32位LX7双核处理器	240MHz	-	40nm

资料来源：各公司官网等，民生证券研究院整理

4 投资建议

4.1 行业投资建议

我们认为，伴随着公开数据预训练scaling law的结束，AI产业叙事开始向云厂商合作伙伴转移。从数据到模型，从训练到推理，从云到端，CSP厂商全面布局，形成了完美的商业闭环。当下，无论是海外的谷歌、亚马逊，还是国内的字节、腾讯，云厂商巨头们开始接力，引领AI产业的下一棒。具体到投资方向，PCB、铜缆、温控、电源等产业链，是国内企业深耕多年，具备优势的环节。伴随着本土云厂商的大力扩产，内需为王的时代也将来临。相比过去，25年的AI产业投资将更重视合作建立、订单落地以及业绩兑现，投资回报也会更为稳健。

建议关注：1、**云端算力：**1) **ASIC：**寒武纪、海光信息、中兴通讯；2) **服务器：**浪潮信息、工业富联、华勤技术、联想集团；3) **AEC：**新易盛、博创科技、瑞可达、兆龙互联、立讯精密；4) **铜连接：**沃尔核材、精达股份；5) **PCB：**生益电子、广合科技、深南电路、威尔高；6) **散热：**申菱环境、英维克、高澜股份；7) **电源：**麦格米特、欧陆通、泰嘉股份。2、**端侧硬件：**1) **品牌：**小米集团、漫步者、亿道信息等；2) **代工：**国光电器、歌尔股份、天键股份、佳禾智能等；3) **数字芯片：**乐鑫科技、恒玄科技、星辰科技、富瀚微、中科蓝讯、炬芯科技、全志科技等；4) **存储芯片：**兆易创新、普冉股份；5) **渠道配镜：**博士眼镜、明月镜片等。

4.2 相关公司梳理

4.2.1 ASIC

中兴通讯：公司在芯片研发领域深耕近30年，不断加强在先进工艺设计、架构创新、封装技术以及核心知识产权等方面的投入，建立了数字化高效开发平台，形成了行业领先的全流程芯片设计能力，专注于ASIC芯片底层技术的研发，并紧跟算网一体化的趋势，围绕“数据、算力、网络”打造高效、环保、智能的全栈算网基础。

4.2.2 AEC

瑞可达：公司专注于连接系统产品的设计开发和制造，系国内知名连接器生产制造商，具备完整的连接器配套产业链。公司正密切关注AI与数据中心产业，未来将重点布局和发展AEC高速组件产品。

兆龙互连：公司是国内领先的数据电缆、专用电缆和连接产品设计与制造的高新技术企业。高速电缆组件是公司未来重点发展方向，公司在服务器内外高速连接

产品上拥有全套自主设计能力，包括 PCBA、线端连接器等，在信号仿真、结构设计和模具设计等领域拥有较深的技术积累。

博创科技：公司产品主要面向电信和数据中心、消费及工业互联领域，已向多家国内外互联网客户批量供货高速铜缆，旗下高性能 800G AEC 系列产品支持 400G/800G 传输速率，涵盖 OSFP/QSFP-DD/QSFP112 封装，最远传输距离达 7 米，较传统 DAC 极大丰富了应用场景。

立讯精密：铜连接是公司业务的核心产品，在数据中心高速互联领域，公司构建了柜内互联、柜间互联、服务器、交换机完整解决方案，并且已协同头部芯片厂商前瞻性为全球主流数据中心及云服务厂商共同制定 800G、1.6T 等下一代高速连接标准。

4.2.3 铜连接

沃尔核材：公司高速铜缆业务由乐庭智联经营，具有丰富的产品开发经验和制程控制经验，通过多年的技术积累，乐庭智联掌握全部重点产品的核心技术，主要产品为 400G/800G 高速通信线，目前 SFP、QSFP、QSFP-DD、SAS、Mini-SAS 等一系列产品取得了众多知名客户的认可，224G 高速通信线产品已稳定批量交付给直接客户。

精达股份：公司全球领先的电磁线制造商之一，主导产品包括铜基电磁线。公司控股子公司恒丰特导主要产品为镀银线、镀镍线、镀锡线、纯银线、纯镍线等产品，具备高性能铜基丝线材制备加工核心技术，开发出具有自主知识产权的成套制备加工技术并成功应用于铜及贵金属丝线材等系列产品。

4.2.4 PCB

生益电子：公司成功开发包括亚马逊在内的多家服务器客户，AI 配套的主板及加速卡项目均已经进入量产阶段，未来新一代产品项目在持续合作开发中。生益电子深耕高精度多层 PCB 与高频高速 PCB 研发，公司多个客户的 AI 产品项目已经成功实现批量，未来新一代的产品项目在持续合作开发中。2024 年前三季度，公司服务器产品营收规模同比实现较大增长，服务器产品占比 42.45%，同比提升 20.87pts。

广合科技：广合科技深耕于高速 PCB 领域，产品主要定位于中高端应用市场，其中服务器用 PCB 产品的收入占比约七成。广合科技的 AI 服务器相关产品的出货占比已超过 25%，且增速显著高于传统服务器，其中 UBB、I/O 板等产品显著起量。公司已通过下游客户认证的英伟达 Blackwell GB200 芯片相关产品，主要聚焦于服务器 CPU 主板。

深南电路：深南电路深耕 PCB 行业 40 年，已成为全球领先的无线基站射频功放 PCB 供应商、内资最大的封装基板供应商。公司前三季度数据中心领域订单同比显著增长，主要得益于 AI 加速卡、EagleStream 平台产品持续放量等产品需

求提升，400G 及以上的高速交换机、光模块产品需求在 AI 相关需求的带动下有所增长，产品结构优化。在新产品预研方面，公司已配合下游客户开展下一代平台产品研发。

威尔高：公司产品覆盖厚铜板、Mini LED 光电板、平面变压器板、光模块等，产品应用于工业控制、显示、汽车电子、新能源、通讯设备、AI 服务器等领域。公司工控业务占比约 40%，目前正在从一次电源向二次电源拓展，**在 AI 领域公司主要包括 DC-DC 相关电源产品 PCB，相关 PCB 层数已做到 30 层。**

4.2.5 散热

申菱环境：公司数据中心液冷产品众多，包括有端到端的全链条解决，核心 CDU 设备，二次侧系统，manifold 管路，一次侧及外部冷源等。新的数据中心产品基地很快就会投产，未来相关业务特别是液冷产品的质量、技术及交付能力都会大幅提升。公司在液冷产品领域参与较早，项目案例较多，与重要客户的技术粘性高，未来相关业务增长趋势较好。

英维克：英维克是业内领先的精密温控节能解决方案和产品提供商，凭借自身掌握的关键自主技术，液冷散热技术已有端到端全链条布局，针对算力设备和数据中心推出的 Coolinside 液冷机柜及全链条液冷解决方案。作为数据中心液冷项目的领军企业，截至今年 4 月，已累计交付 900MW 液冷项目，技术实力与市场地位稳固，项目执行经验丰富。

高澜股份：公司作为国内最早聚焦电力电子热管理技术创新和产业化应用的企业之一，拥有行业领先的技术，热管理业务主要产品达到国内领先或国际先进水平，部分产品达到国际领先水平。公司的液冷解决方案以冷板式为主，液冷产品的相关客户包含字节跳动、阿里巴巴、腾讯、万国数据、浪潮等企业，公司积极布局海外市场，热管理产品已在全球多个国家和地区投入使用，应用场景涉及大功率电力电子、数据中心等领域。公司产品已取得 CE、ETL、ASMEU、ROHS 等认证，并已直接或间接应用于欧洲、美国等海外项目。

4.2.6 电源

麦格米特：公司是行业领先的电源解决方案提供商，具备业界领先的高功率高效率网络电源的技术水平及产品研发与供应能力。2024年10月17日，公司宣布与英伟达展开合作，将为NVIDIA MGX™平台和GB200系统提供先进的电源解决方案。

欧陆通：公司深耕电源领域多年，配置有全功能、全方位的研发与产品综合性实验室，产品技术参数均可实现自主设计、检测、实验，保证了研发速度与品质标准，主要产品包括电源适配器、数据中心电源和其他电源等。

表8：重点关注个股

证券代码	证券简称	股价 (元)	EPS			PE			评级
			2024E	2025E	2026E	2024E	2025E	2026E	
688041.SH	海光信息	155.05	0.84	1.12	1.46	185	138	106	推荐
688800.SH	瑞可达	56.19	1.04	1.38	1.84	54	41	31	推荐
002851.SZ	麦格米特	60.72	1.13	1.53	1.95	54	40	31	/
300870.SZ	欧陆通	111.99	2.11	2.87	3.91	53	39	29	/
300499.SZ	高澜股份	22.16	0.10	0.27	0.43	218	83	52	/
301018.SZ	申菱环境	38.98	0.77	1.02	1.24	51	38	31	/
002130.SZ	沃尔核材	29.22	0.73	0.95	1.14	40	31	26	/
600577.SH	精达股份	8.02	0.26	0.34	0.41	31	23	20	/
688629.SH	华丰科技	38.89	0.20	0.46	0.63	196	84	62	/

资料来源：，民生证券研究院预测；
(注：股价为 2024 年 12 月 27 日收盘价；未覆盖公司数据采用 wind 一致预期)

5 风险提示

1) 大模型发展不及预期。大模型的发展可能会遇到阶段性瓶颈，若用于训练的公共数据消耗殆尽，则大模型的迭代可能会放缓，难以吸引用户付费进行体验。

2) CSP 自研算力不及预期。相较于英伟达，CSP 自研 AI 算力起步更晚，软件生态相对不足，如果 CSP 自研 AI 算力不及预期，则可能对产业链落地节奏及合作伙伴业绩带来不利影响。

3) AI 终端销量不及预期。电子产品销量具有不确定性，若大模型创新不足以吸引用户付费体验，AI 终端的下游需求不佳，将对板块内公司业绩造成不利影响。

插图目录

图 1: AI 大模型对算力的需求超过摩尔定律	3
图 2: Scaling Law 的三要素: 算力、数据量、参数规模	4
图 3: 2018-2028 年全球数据量	4
图 4: 大模型 Scaling Law 将在 2028 年左右开始显著放缓	4
图 5: 低精度的训练数据增多可能反而对模型性能造成损害	5
图 6: 使用更高精度的数据将减小因数据质量不佳而对模型性能造成的损害	5
图 7: 豆包大模型使用量快速增长	6
图 8: 推理对模型重要性提升	7
图 9: 谷歌 TPU	7
图 10: AI 发展历程复盘	8
图 11: 1Q20-4Q24 北美云商资本开支 (亿美元)	10
图 12: 2021-2024E 北美云商资本开支 (亿美元)	10
图 13: 北美云商云业务收入 (亿美元) 及同比增速	10
图 14: 国内云商云业务收入 (亿美元) 及同比增速	10
图 15: 2022-2024 年全球 AI 芯片竞争格局	11
图 16: Trainium2 芯片	13
图 17: Trainium2 内部组件	13
图 18: 博通两大核心客户 AI 芯片战略规划	15
图 19: 博通两大核心客户 AI 芯片战略规划	15
图 20: 传统服务器 PCB 应用	16
图 21: DGX H100 AI 服务器中 OAM 和 UBB (绿色)	16
图 22: GB200 服务器 PCB 方案发生变化	17
图 23: 搭载有谷歌第六代 TPU 芯片 Trillium 的 PCB	17
图 24: 亚马逊 Trainium2 Compute Tray	18
图 26: 英伟达 DGX H100 GPU Tray	18
图 27: 英伟达 GB200 NVL 72 机柜架构	19
图 28: GB200 NVL72 架构的通信解决方案	20
图 29: Rack 内部铜互联解决方案	20
图 30: Rack 之间的铜互联解决方案	20
图 31: 全球 DAC、AEC、AOC 市场规模 (十亿美元)	21
图 32: DAC、AEC、AOC 功耗对比	22
图 33: DAC、AEC、AOC 成本对比	22
图 34: Trainium2 Rack	23
图 35: Trainium2-Ultra Rack	23
图 36: 风冷及液冷对比	24
图 37: 冷板式液冷系统示意图	25
图 38: 浸没式液冷原理示意图	25
图 39: AI 带动服务器机柜功率及功率密度持续提升	26
图 40: 机架式服务器系统及电源框图	27
图 41: AI 服务器采用三级电源解决方案	27
图 42: 国内各大模型 2024 年 (截至 11 月 15 日) 广告投放投入	29
图 43: 字节跳动 2024 年硬件布局	29
图 44: 豆包通用模型 pro 能力全面提升	31
图 45: Trainium2-Ultra Rack	31
图 46: 火山引擎时间安排表	31
图 47: AI 终端广阔空间	32
图 48: AI+硬件持续落地	33

表格目录

重点公司盈利预测、估值与评级	1
表 1: 英伟达客户在加速卡领域的布局情况	12
表 2: Trainum2 芯片参数	13
表 3: DAC、AEC、AOC 性能对比	22
表 4: 豆包大模型推理需求测算	30
表 5: 国产算力+端侧产业链公司	31
表 6: 数字芯片公司合作伙伴梳理	34
表 7: 主要算力芯片参数	35
表 8: 重点关注个股	39

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。本报告清晰准确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
公司评级	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
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