电子AI+系列专题报告（六）

DeepSeek重塑开源大模型生态，AI应用爆发持续推升算力需求

行业研究 · 行业专题

电子

投资评级：优于大市（维持）

证券分析师：胡剑
021-60893306
hujian1@guosen.com.cn
S0980521080001

证券分析师：胡慧
021-60871321
huhui2@guosen.com.cn
S0980521080002

证券分析师：叶子
0755-81982153
yezi3@guosen.com.cn
S0980522100003

证券分析师：张大为
021-61761072
zhangdawei1@guosen.com.cn
S0980524100002

证券分析师：詹浏洋
010-88005307
zhanliuyang@guosen.com.cn
S0980524060001

DeepSeek重塑开源大模型生态，AI应用爆发持续推升算力需求

国信证券 GUOSEN SECURITIES

- **DeepSeek发展突飞猛进，领跑开源大模型技术与生态，DeepSeek模型已成为全球现象级模型。**DeepSeek（深度求索）公司成立于2023年7月，是一家致力于实现通用人工智能（AGI）的创新型科技公司。2024年12月，DeepSeek-V3发布，性能对齐海外领军闭源模型。据官方技术论文披露，V3模型的总训练成本为557.6万美元，对比GPT-4o等模型的训练成本约为1亿美元。2025年1月，DeepSeek-R1发布，性能对标OpenAI-o1正式版。在数学、代码、自然语言推理等任务上，性能比肩OpenAI-o1正式版。2月1日消息，据彭博社报道，DeepSeek的人工智能助手在140个市场下载次数最多的移动应用程序排行榜上名列前茅。国外大型科技公司如微软、英伟达、亚马逊等已先后上线部署支持用户访问DeepSeek-R1模型。2月1日，华为云官方发布消息，硅基流动和华为云团队联合首发并上线基于华为云昇腾云服务的DeepSeekR1/V3推理服务。

- **DeepSeek通过MLA和DeepSeekMoE实现高效的推理和低成本训练，构建DualPipe算法和混合精度训练优化计算与通信负载；通过（分阶段）强化学习实现性能突破。**多头潜在注意力（MLA）通过低秩联合压缩技术，大幅削减了注意力键（keys）和值（values）的存储空间，显著降低了内存需求。DeepSeekMoE架构采用了更为精细粒度的专家设置，能够更加灵活且高效地调配资源，进一步提升了整体的运行效率和表现。DeepSeek模型对跨节点的全对全通信机制进行优化，充分利用InfiniBand和NVLink提供的高带宽。创新性提出了DualPipe算法，通过优化计算与通信的重叠，有效减少了流水线中的空闲时间。采用FP8混合精度训练技术，不仅极大地加快了训练速度，还大幅降低了GPU内存的消耗。DeepSeek-R1-Zero通过强化学习架构创新实现突破性性能，核心技术创新体现在训练效能优化策略、双维度评价体系、结构化训练范式三个维度。DeepSeek-R1采用分阶段强化学习架构演进，包括冷启动阶段、面向推理的强化学习、拒绝采样与监督式微调、全场景强化学习等。

- **AI应用爆发在即，算力需求持续攀升，关注ASIC及服务器产业链。**Scaling Law与“涌现”能力是大模型训练遵循的重要法则，随着ChatGPT引领全球AI浪潮，国内外科技公司纷纷发布AI大模型，截至24年7月，全球AI大模型数量约1328个（其中美国位居第一位，占比44%；中国位居第二位，占比36%），模型的迭代加速、竞争加剧。同时，AI模型向多模态全方位转变，AI应用百花齐放，企业主动拥抱AI应用市场。因此，模型数量、模型参数、数据总量的持续增长及AI应用需求推动全球算力爆发式增长。在英伟达GPU随着架构的不断演进及算力的成倍增长，于AI大模型训练中得到广泛运用的同时，为了满足CSP客户更高性能和更好功能的需求，定制化芯片ASIC的需求持续提升，牧本钟摆从标准化逐渐摆向定制化。与之相应的算力基础设施持续建设和升级，促使国内外云服务商资本开支持续高速增长，带来AI服务器市场规模大幅提升，预计到26年全球AI服务器出货量将达到237万台，对应2023-2026年CAGR为26%。

- **重点推荐组合：**中芯国际、翱捷科技、德明利、工业富联、沪电股份、联想集团、国芯科技、澜起科技、芯原股份、龙芯中科、东山精密、景旺电子。

- **风险提示：**宏观AI应用推广不及预期、AI投资规模低于预期、AI服务器渗透率提升低于预期、AI监管政策收紧。

目录

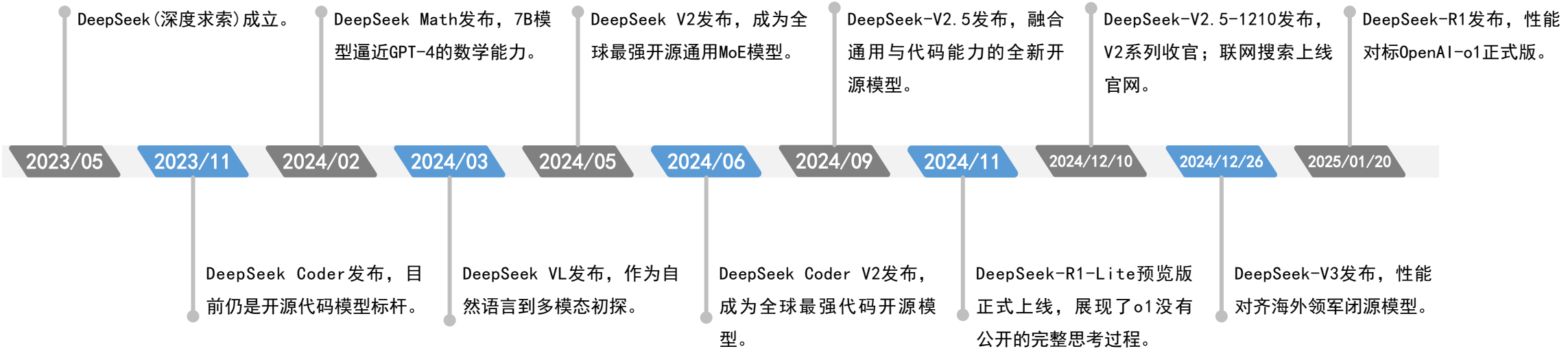
DeepSeek发展突飞猛进，领跑开源大模型技术与生态

DeepSeek成立不到两年颠覆开源大模型格局，性能对标海外

国信证券 GUOSEN SECURITIES

- DeepSeek(深度求索)公司成立于2023年5月，是一家致力于实现AGI(Artificial General Intelligence，通用人工智能)的创新型科技公司，专注于开发先进的大语言模型和相关技术。DeepSeek由知名量化资管巨头幻方量化创立，幻方量化创始人梁文峰在量化投资和高性能计算领域具有深厚的背景和丰富的经验。
- **2024年5月，DeepSeek-V2发布，成为全球最强开源通用MoE模型。**DeepSeek独创Attention结构MLA(一种新的多头潜在注意力机制)、稀疏结构DeepSeek-MoE在大模型竞技场(LMSYS)位列全球开源模型第一名，依靠创新结构，将推理成本降低近百倍。
- **2024年12月，DeepSeek-V3发布，性能对齐海外领军闭源模型。**该模型在多项评测集上超越了阿里Qwen2.5-72B、Meta的Llama-3.1-405B等其他开源模型，并逼近GPT-4o、Claude-3.5-Sonnet等顶尖闭源模型。据官方技术论文披露，V3模型的总训练成本为557.6万美元，对比GPT-4o等模型的训练成本约为1亿美元。
- **2025年1月，DeepSeek-R1发布，性能对标OpenAI-o1正式版。**DeepSeek-R1在后训练阶段大规模使用了强化学习技术，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，性能比肩OpenAI-o1正式版。同时DeepSeek开源R1推理模型，允许所有人在遵循MIT License的情况下，蒸馏R1训练其他模型。

图：DeepSeek模型迭代与发展历史沿革



资料来源：DeepSeek官网，国信证券经济研究所整理

DeepSeek模型已成为全球现象级大模型

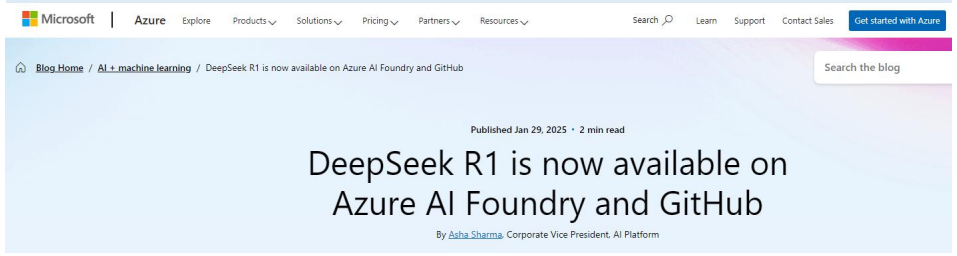
国信证券 GUOSEN SECURITIES

- **DeepSeek下载量占据140多个国家榜首。**2月1日消息，据彭博社报道，DeepSeek的人工智能助手在140个市场下载次数最多的移动应用程序排行榜上名列前茅，其中印度占据了新用户的最大比例。据Appfigures数据（不包括中国的第三方应用商店），这款推理人工智能聊天机器人于1月26日升至苹果公司应用商店的榜首，此后一直占据全球第一的位置。

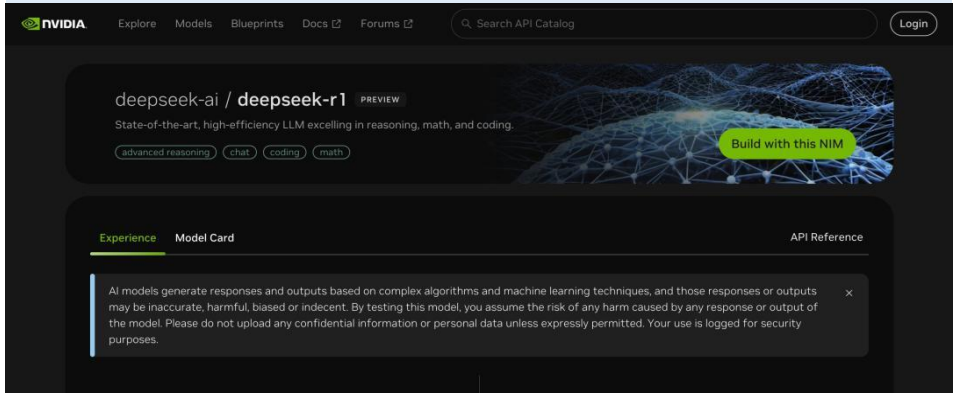
- **国外大型科技公司已上线部署支持用户访问DeepSeek-R1模型。**1月30日，微软宣布DeepSeek-R1模型已在Azure AI Foundry和GitHub上提供。1月31日，英伟达宣布DeepSeek-R1模型已作为NVIDIA NIM微服务预览版在英伟达面向开发者的网站上发布；同日亚马逊宣布，客户现已可以在Amazon Bedrock和Amazon SageMaker AI中部署DeepSeek-R1模型。

- **硅基流动和华为云宣布上线DeepSeekR1/V3推理服务。**2月1日，华为云官方发布消息，硅基流动和华为云团队联合首发并上线基于华为云昇腾云服务的DeepSeekR1/V3推理服务。该服务具备以下特点：1）得益于自研推理加速引擎加持，硅基流动和华为云昇腾云服务支持部署的DeepSeek模型可获得持平全球高端GPU部署模型的效果。2）提供稳定的、生产级服务能力，让模型能够在大规模生产环境中稳定运行，并满足业务商用部署需求。华为云昇腾云服务可以提供澎湃、弹性、充足的算力。

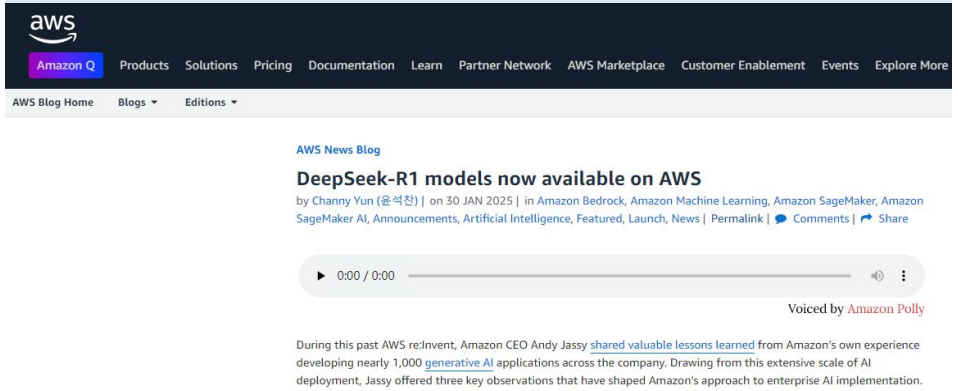
图：微软宣布支持访问DeepSeek-R1模型



图：英伟达宣布支持访问DeepSeek-R1模型



图：亚马逊宣布支持访问DeepSeek-R1模型



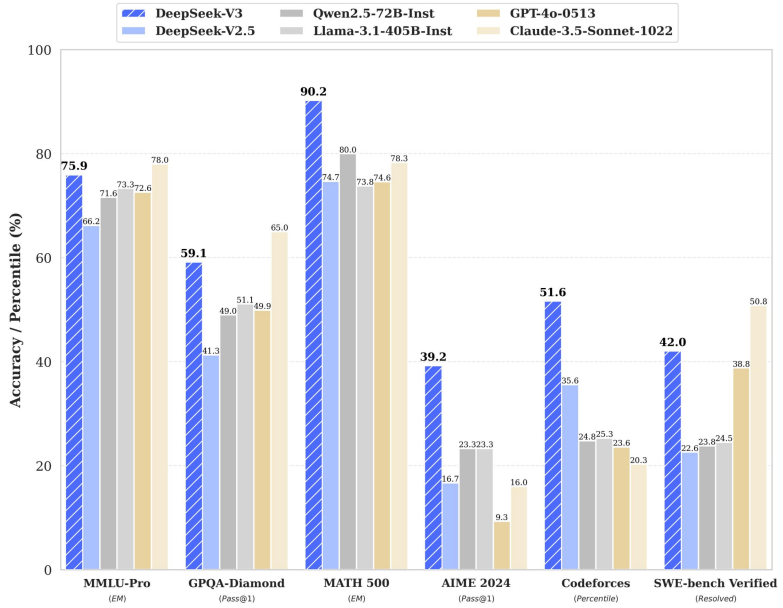
资料来源：各公司官网，国信证券经济研究所整理

DeepSeek-V3发布，性能对齐海外领军闭源模型

国信证券 GUOSEN SECURITIES

- **2024年12月26日，全新系列模型DeepSeek-V3首个版本上线并同步开源。**DeepSeek-V3为自研MoE模型，共有671B参数，每个token激活37B，在14.8T token上进行预训练。DeepSeek-V3多项评测成绩超越了Qwen2.5-72B和Llama-3.1-405B等其他开源模型，并在性能上和世界顶尖的闭源模型GPT-4o及Claude-3.5-Sonnet不分伯仲。
- **DeepSeek-V3模型生成速度提升至3倍。**通过算法和工程上的创新，DeepSeek-V3的生成吐字速度从20TPS大幅提高至60TPS，相比V2.5模型实现了3倍的提升，能够为用户带来更加迅速流畅的使用体验。
- **DeepSeek-V3模型具有更优的模型性能/价格比例。**随着性能更强、速度更快的DeepSeek-V3更新上线，模型API服务定价调整为每百万输入tokens 0.5元(缓存命中)/2元(缓存未命中)，每百万输出tokens8元。相比于其他模型性能和定价，该模型具有更优的模型性能/价格比例。

图：DeepSeek-V3等模型执行不同推理任务性能对比



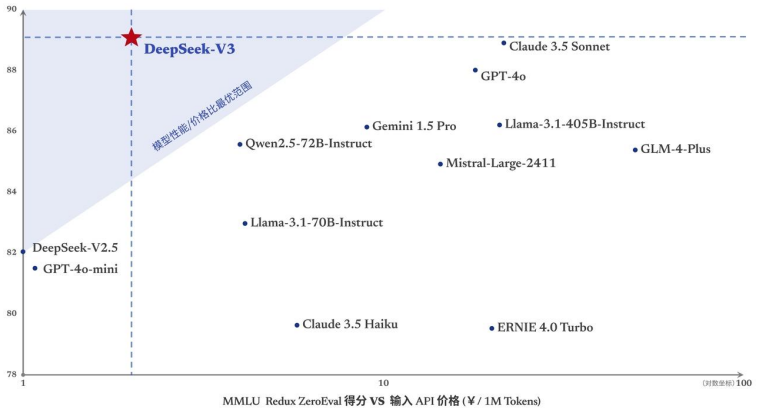
资料来源：DeepSeek官网，国信证券经济研究所整理

图：DeepSeek-V3等模型多项评测成绩对比

测试集		DeepSeek-V3	Qwen2.5 72B-Inst.	Llama3.1 405B-Inst.	Claude-3.5-Sonnet-1022	GPT-4o 0513
模型架构		MoE	Dense	Dense	-	-
# 激活参数		37B	72B	405B	-	-
# 总参数		671B	72B	405B	-	-
英文	MMLU (EM)	88.5	85.3	88.6	88.3	87.2
	MMLU-Redux (EM)	89.1	85.6	86.2	88.9	88
	MMLU-Pro (EM)	75.9	71.6	73.3	78	72.6
	DROP (3-shot F1)	91.6	76.7	88.7	88.3	83.7
	IF-Eval (Prompt Strict)	86.1	84.1	86	86.5	84.3
	GPQA-Diamond (Pass@1)	59.1	49	51.1	65	49.9
	SimpleQA (Correct)	24.9	9.1	17.1	28.4	38.2
	FRAMES (Acc.)	73.3	69.8	70	72.5	80.5
	LongBench v2 (Acc.)	48.7	39.4	36.1	41	48.1
代码	HumanEval-Mul (Pass@1)	82.6	77.3	77.2	81.7	80.5
	LiveCodeBench(Pass@1-COT)	40.5	31.1	28.4	36.3	33.4
	LiveCodeBench (Pass@1)	37.6	28.7	30.1	32.8	34.2
	Codeforces (Percentile)	51.6	24.8	25.3	20.3	23.6
	SWE Verified (Resolved)	42	23.8	24.5	50.8	38.8
	Aider-Edit (Acc.)	79.7	65.4	63.9	84.2	72.9
	Aider-Polyglot (Acc.)	49.6	7.6	5.8	45.3	16
数学	AIME 2024 (Pass@1)	39.2	23.3	23.3	16	9.3
	MATH-500 (EM)	90.2	80	73.8	78.3	74.6
	CNMO 2024 (Pass@1)	43.2	15.9	6.8	13.1	10.8
中文	CLUEWSC (EM)	90.9	91.4	84.7	85.4	87.9
	C-Eval (EM)	86.5	86.1	61.5	76.7	76
	C-SimpleQA (Correct)	64.1	48.4	50.4	51.3	59.3

资料来源：DeepSeek官网，国信证券经济研究所整理

图：DeepSeek-V3具有更优的模型性能/价格比例



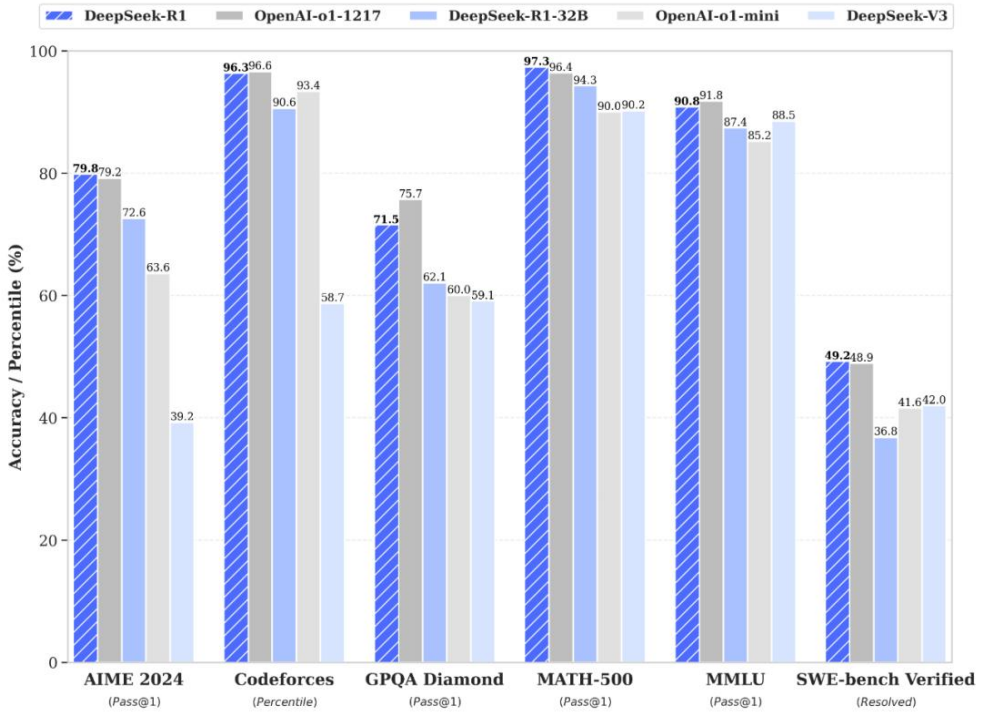
资料来源：DeepSeek官网，国信证券经济研究所整理

DeepSeek-R1发布，性能对标OpenAI-o1正式版

国信证券 GUOSEN SECURITIES

- **2025年1月20日，DeepSeek-R1正式发布，并同步开源模型权重，性能对齐OpenAI-o1正式版。**DeepSeek-R1在后训练阶段大规模使用了强化学习技术，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，性能比肩OpenAI-o1正式版。
- DeepSeek-R1遵循MIT License，允许用户通过蒸馏技术借助R1训练其他模型；同时上线API，对用户开放思维链输出；DeepSeek官网与App同步更新上线，用户打开“深度思考”模式，即可调用最新版DeepSeek-R1完成各类推理任务。
- **DeepSeek蒸馏小模型超越OpenAIo1-mini。**在开源DeepSeek-R1-Zero和DeepSeek-R1两个660B模型的同时，通过DeepSeek-R1的输出，蒸馏了6个小模型开源给社区，其中32B和70B模型在多项能力上实现了对标OpenAIo1-mini的效果。

图：DeepSeek-R1等模型执行不同推理任务性能对比



资料来源：DeepSeek官网，国信证券经济研究所整理

图：DeepSeek蒸馏小模型等执行不同推理任务性能对比

	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官网，国信证券经济研究所整理

DeepSeek-V3通过MLA和DeepSeekMoE实现高效的推理和低成本训练

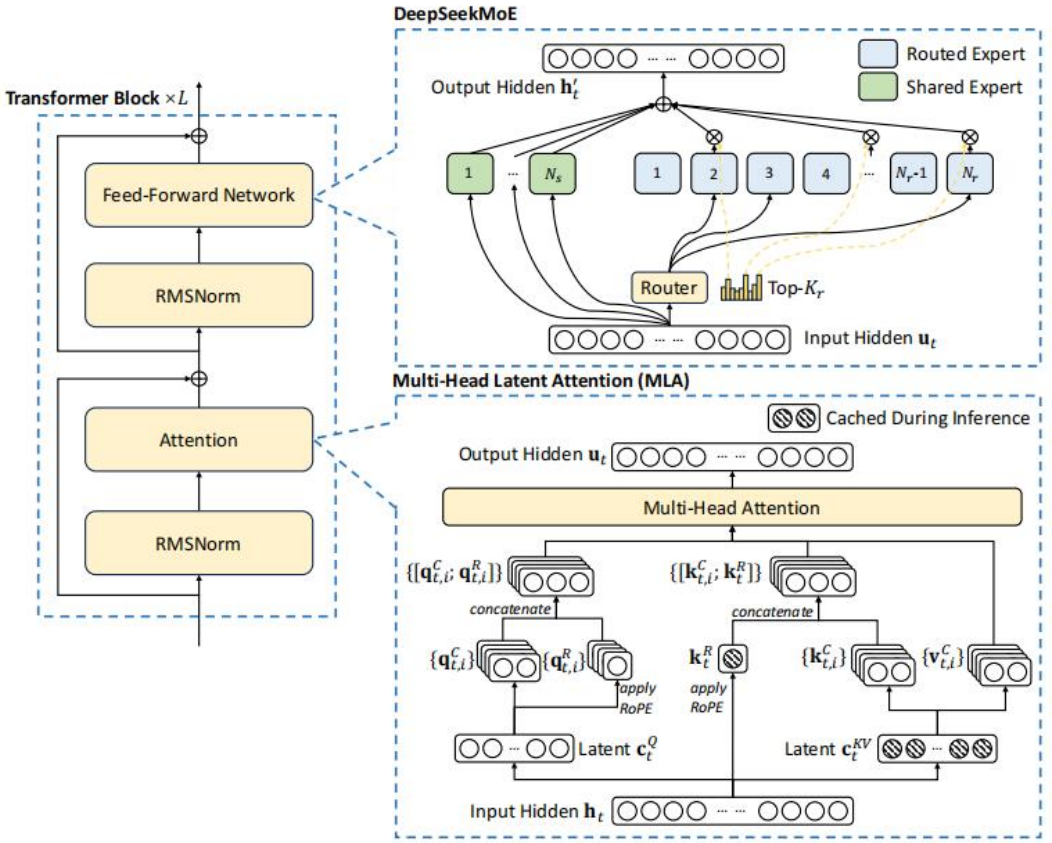
国信证券 GUOSEN SECURITIES

- **DeepSeek-V3以Transformer框架为基石，创新性地融入多头潜在注意力(Multi-head Latent Attention，MLA)和DeepSeekMoE架构。**这一设计在维持模型高性能的同时，极大地提升了训练与推理的效率。

- **多头潜在注意力(MLA)：**在传统的注意力机制中，推理期间的键值(Key-Value，即KV)缓存往往占用大量资源。而MLA则另辟蹊径，通过低秩联合压缩技术，大幅削减了注意力键(keys)和值(values)的存储空间。在生成过程中，仅需缓存压缩后的潜在向量，这一举措显著降低了内存需求，但在性能上与标准多头注意力(Multi-head Attention，MHA)相比毫不逊色，有力地保障了模型运行的流畅性。

- **DeepSeekMoE架构：**该架构采用了更为精细粒度的专家设置，还特别将部分专家设定为共享专家。在每一个MoE层中，都由共享专家和路由专家协同构成。其中，共享专家负责处理所有token的输入信息，为模型提供基础的处理支撑；而路由专家则依据每个token与专家之间的亲和度分数(这一分数通过sigmoid函数计算得出，即token-to-expert affinity)来决定是否被激活。这种独特的设计，使得模型在处理不同类型的输入时，能够更加灵活且高效地调配资源，进一步提升了整体的运行效率和表现。

图：DeepSeek-V3模型技术架构



资料来源：DeepSeek-V3技术报告，国信证券经济研究所整理

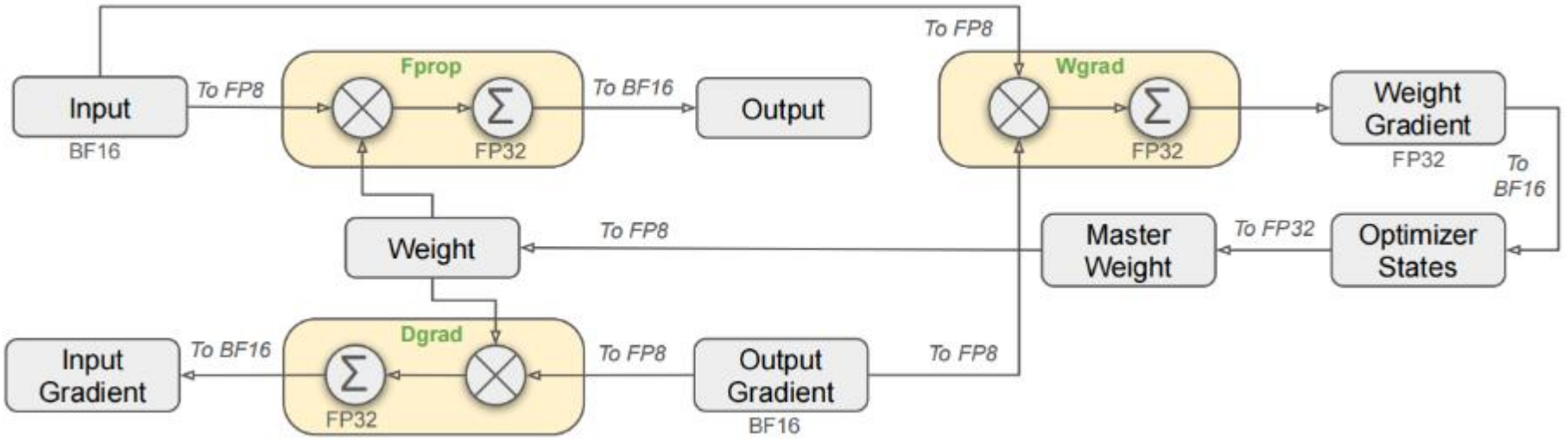
DeepSeek-V3构建DualPipe算法和混合精度训练优化计算与通信负载

- **对跨节点的全对全通信机制进行优化，充分利用InfiniBand和NVLink提供的高带宽。**DeepSeek-V3模型在拥有2048个NVIDIA H800 GPU的大规模集群上进行训练，每个节点配置了8个GPU，并通过NVLink与NVSwitch实现内部高速互联；不同节点间的高效通信则依赖于InfiniBand（IB）网络。

- **创新性提出了DualPipe算法，通过优化计算与通信的重叠，有效减少了流水线中的空闲时间。**对于DeepSeek-V3而言，由于跨节点专家并行引入的通信开销导致计算与通信的比例接近1:1，因此提出DualPipe（双向管道并行）算法，采用一种新的双向流水线方法，在独立的前向和后向处理块中实现了计算与通信的重叠，从而加速模型的训练过程并降低了气泡效应。为了确保DualPipe的性能最优，定制设计了高效的跨节点全对全通信核心，包括优化的调度和组合策略，减少用于通信的流式多处理器（SMs）资源占用，并通过调优PTX指令集和自动调整通信数据块大小，显著减少了L2缓存的使用及对其他SMs的干扰。

- **采用FP8混合精度训练技术，不仅极大地加快了训练速度，还大幅降低了GPU内存的消耗。**基于低精度训练领域的成熟经验，开发人员构建了一个适用于FP8训练的混合精度框架，其中大部分计算密集型任务以FP8精度执行，而关键操作则保持原有精度，以确保数值稳定性和训练效率之间的平衡。结合FP8训练框架，能够将缓存激活值和优化器状态压缩至低精度格式，进一步减少了内存占用和通信负载。

图：采用FP8数据格式的整体混合精度框架



资料来源：DeepSeek-V3技术报告，国信证券经济研究所整理

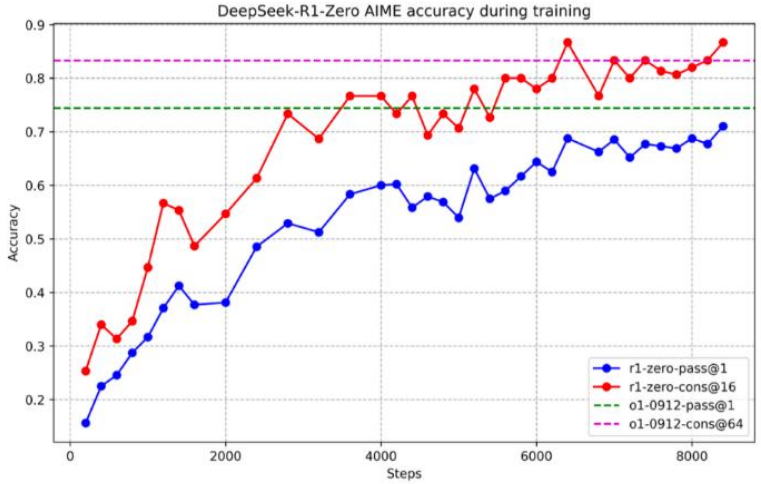
DeepSeek-R1(-Zero)通过(分阶段)强化学习实现性能突破

国信证券 GUOSEN SECURITIES

- **DeepSeek-R1-Zero：通过强化学习架构创新实现突破性性能。**该模型突破性地采用纯强化学习(RL)方法，未经过传统监督式微调(SFT)即达成卓越性能表现，在特定任务基准测试中实现对OpenAI-o1的超越。其核心技术创新体现在三个维度：1）**训练效能优化策略。**创新性采用GRPO(群体相对策略优化)算法，该技术继承自DeepSeek-V2的RLHF(人类反馈强化学习)研发成果。与传统方法相比，GRPO通过群体反馈数据分析替代独立评估模型，有效降低计算资源消耗。这种优化策略无需依赖与策略模型规模匹配的独立评估模型，通过动态基线估计显著提升训练效率。2）**双维度评价体系。**建立“准确性验证+格式规范”的复合奖励机制：前者通过数学符号解析与代码编译测试进行精确度验证，后者要求模型将推理过程严格置于<think>结构化标签内。这种双重设计既保障了技术问题求解的严谨性，又确保了输出内容的可解析性，为自动化评估提供标准化接口。3）**结构化训练范式。**研发团队设计了标准化指令模板，通过分离推理过程与最终结论的结构化输出要求，既保证了知识表达的清晰度，又保留了内容创作的自主性。该模板仅规范输出框架，避免对具体解题方法或思维路径进行预设性限制。

- **DeepSeek-R1：分阶段强化学习架构演进。**为克服Zero版本存在的可读性差、语言混淆的问题并提升结果校准能力，该迭代版本采用多阶段强化学习策略：1）**冷启动阶段：**通过少量高质量思维链(CoT)示范数据进行模型初始化，有效缓解基础模型在初始训练阶段的波动性。2）**面向推理的强化学习。**和DeepSeek-R1-Zero方式相同，但引入了语言一致性奖励，对推理密集型任务进行特别优化。3）**拒绝采样与监督式微调。**使用已训练的RL模型来生成新的训练数据，通过构建推理数据和非推理数据提升模型的通用能力。4）**全场景强化学习。**为了同时平衡推理能力和通用能力，将不同类型的奖励机制有机结合，再次进行强化学习。

图：DeepSeek-R1-Zero训练中针对AIME正确率持续提高



资料来源：DeepSeek-R1技术报告，国信证券经济研究所整理

图：DeepSeek-R1-Zero模型结构化训练模版

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>. User: prompt. Assistant:

Table 1 | Template for DeepSeek-R1-Zero. prompt will be replaced with the specific reasoning question during training.

资料来源：DeepSeek-R1技术报告，国信证券经济研究所整理

采用硬盘缓存技术大幅降低延迟和成本

国信证券 GUOSEN SECURITIES

- **DeepSeek API以其开创性的硬盘缓存技术，实现了价格的指数级下降。**

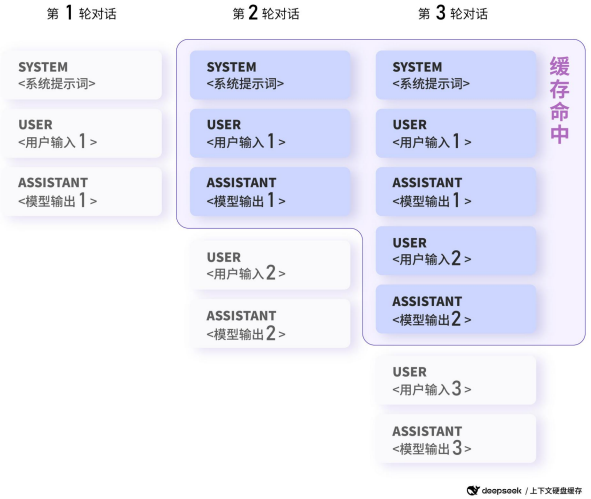
在大模型API的实际运用场景中，用户输入存在较高比例的重复内容。例如，用户输入的提示词(prompt)常常包含重复引用部分；在多轮对话里，每一轮都需重复输入前几轮的内容。针对这些情况，DeepSeek引入上下文硬盘缓存技术，将预估未来可能复用的内容，缓存至分布式硬盘阵列之中。一旦出现重复输入，重复部分直接从缓存读取，无需重新计算。这一技术不仅有效缩短了服务延迟，还极大地降低了最终的使用成本。

- **1）降低服务延迟：**对于输入内容长且重复部分多的请求，API服务的首token延迟会大幅降低。以128K输入且大部分内容重复的请求为例，经实际测试，首token延迟从原本的13秒锐减至500毫秒。
- **2）削减整体费用：**最高能够节省90%的费用(前提是针对缓存特性进行优化)。即便不做任何优化，按照过往使用数据统计，用户整体节省的费用也能超过50%。并且，缓存所占用的存储无需额外付费。
- **3）保障缓存安全：**在设计缓存系统时，DeepSeek已全面考量各类潜在安全问题。每个用户的缓存相互独立，在逻辑层面彼此不可见，从底层架构筑牢用户数据的安全与隐私防线。长时间未使用的缓存会自动清空，不会长期留存，也不会被挪作他用。

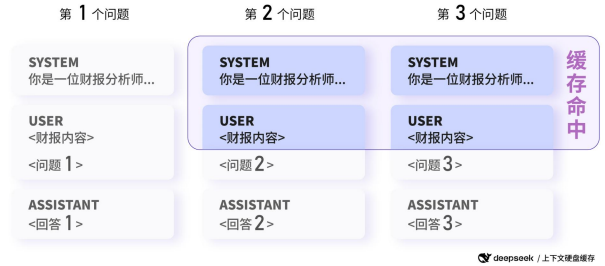
- **DeepSeek可能是全球首家在API服务中大规模应用硬盘缓存的大模型厂商。**

这一成果得益于DeepSeek-V2提出的MLA结构，该结构在提升模型效果的同时，极大地压缩了上下文KVCache的大小，使得存储所需的传输带宽和存储容量大幅降低，进而能够将缓存置于低成本的硬盘之上。

图：多轮对话场景，下一轮对话会命中上一轮对话生成的上下文缓存



图：数据分析场景，后续具有相同前缀的请求会命中上下文缓存

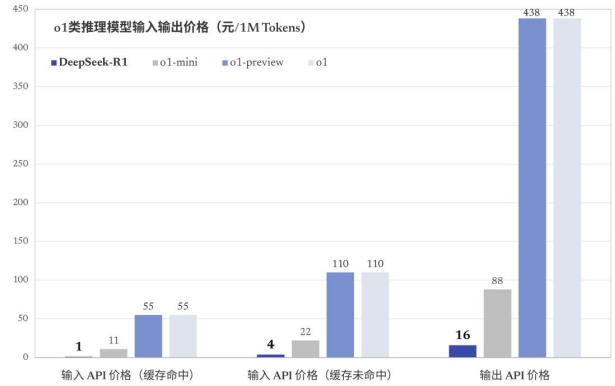


表：DeepSeek-R1 API服务定价

模型	上下文长度	最大思维链长度	最大输出长度	百万tokens输入价格(缓存命中)	百万tokens输入价格(缓存未命中)	百万tokens输出价格
DeepSeek-V3	64K	-	8K	0.5元	2元	8元
DeepSeek-R1	64K	32K	8K	1元	4元	16元

注1：表格中所列模型价格以“百万tokens”为单位。Token是模型用来表示自然语言文本的的最小单位，可以是一个词、一个数字或一个标点符号等。公司将根据模型输入和输出的总token数进行计量计费。
注2：思维链为DeepSeek-R1模型在给出正式回答之前的思考过程。
注3：如未指定max_tokens，默认最大输出长度为4K。可调整该参数以支持更长的输出。
注4：表格中展示了DeepSeek-V3模型优惠前的价格。即日起至北京时间2025-02-08 24:00，所有用户均可享受DeepSeek-V3 API的价格优惠。 在此之后，模型价格将恢复至原价。DeepSeek-R1不参与优惠。
注5：DeepSeek-R1的输出token数包含了思维链和最终答案的所有token，其计价相同。

图：DeepSeek-R1等模型输入输出价格对比



资料来源：DeepSeek官网，国信证券经济研究所整理

AI应用爆发在即，算力需求持续攀升，关注ASIC及服务器产业链

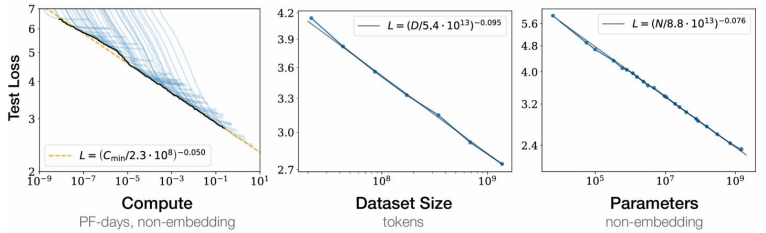
Scaling Law与“涌现”能力：大模型训练遵循的重要法则

国信证券 GUOSEN SECURITIES

- **Scaling Law：模型效果随模型规模指数增加而线性提高。**据OpenAI发布的论文《Scaling laws for neural language models》，模型性能极大依赖训练规模，模型参数、数据集大小以及用于训练的计算量增加可以达到减少模型损失，增加大模型性能的效果。

- **“涌现”能力：随着训练规模不断增大，大模型将产生质变。**据《Emergent Abilities of Large Language Models》，随着模型规模的扩大，语言模型表现出的新的、不可预测的能力。这些新能力在中小模型上线性放大都得不到线性的增长，但在模型规模突破一定阈值时突然出现。“涌现”能力反映了系统行为质的变化，这种变化不能简单地通过观察或分析较小规模模型的性能来预测。

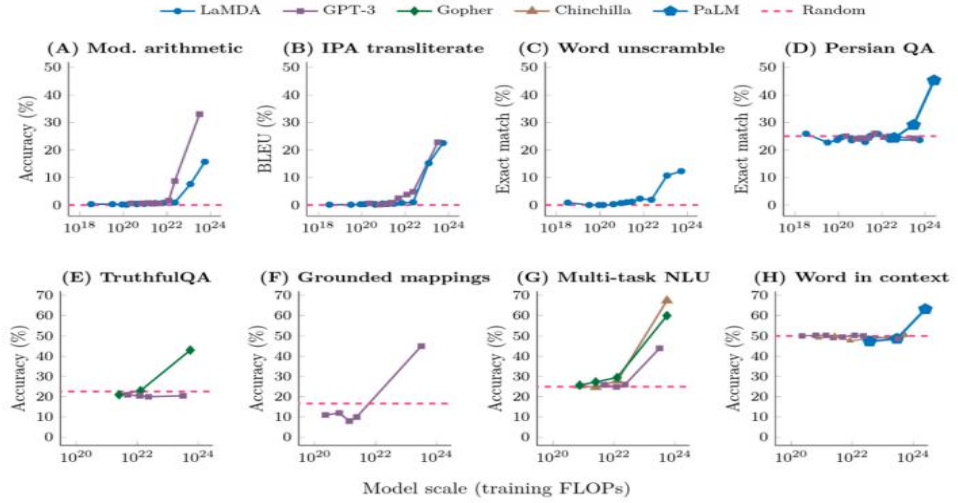
图：模型规模的指数提升线性提高模型性能



资料来源：Jared等著-《Scaling Laws for Neural Language Models》-Arxiv（2020）-P3，国信证券经济研究所整理

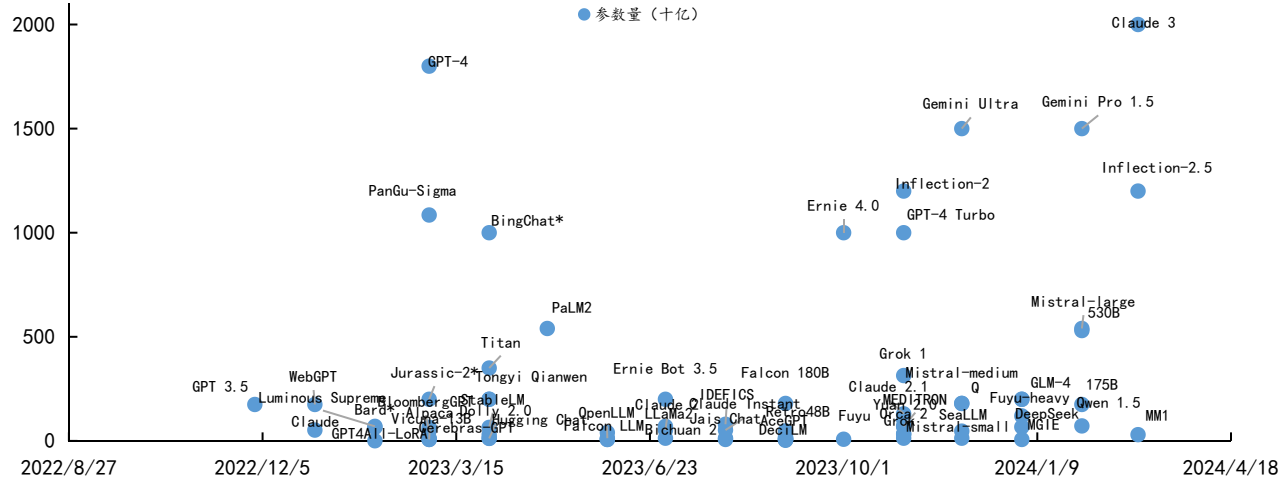
●自1956年计算机专家约翰·麦卡锡提出“人工智能”概念以来，在过去的近70年时间里，行业经历了以CNN为代表的传统神经网络模型、以Transformer为代表的全新神经网络模型、以GPT为代表的预训练大模型这三个时代的进阶，在“算力芯片、存储芯片”等硬件技术持续演进的支撑下，**伴随模型参数规模超越千亿级，近年来人工智能技术得以“涌现”出更加强大的理解、推理、联想能力。**

图：大模型随参数规模增加所体现的“涌现”能力



资料来源：Jared等《Scaling Laws for Neural Language Models》，国信证券经济研究所整理

图：大模型参数量近年来迅速扩容



资料来源：Information is Beautiful，国信证券经济研究所整理

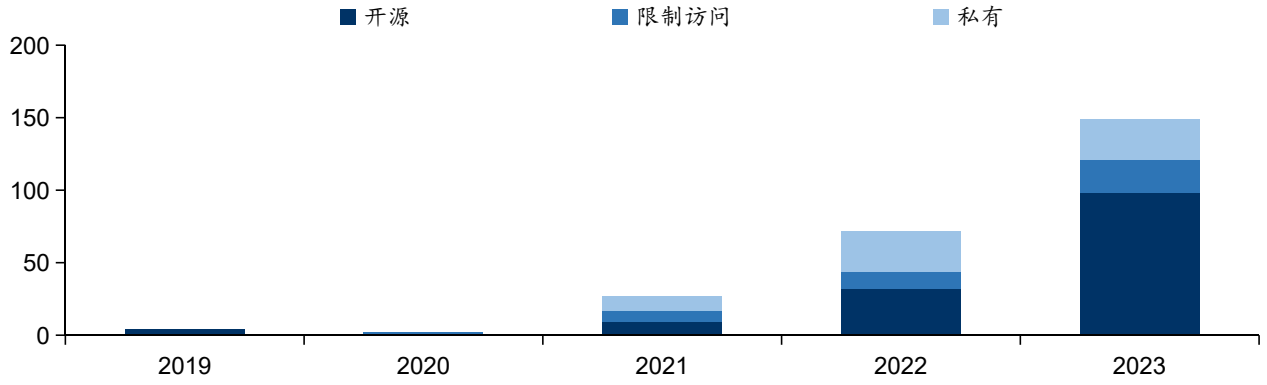
Scaling Law与“涌现”能力：大模型训练遵循的重要法则

国信证券 GUOSEN SECURITIES

- **海内外科技公司纷纷发布AI大模型，模型的更新迭代和竞争加剧。**据中国信通院数据，截至2024年7月，全球AI大模型数量约1328个（包含同一企业、同一模型的不同参数版本），其中美国AI大模型数量位居第一位，占比44%，代表性模型包括OpenAI的GPT、Anthropic的Claude、Meta的Llama、Google的Gemini等；中国AI大模型数量位居第二位，占比36%，代表性模型包括阿里的通义千问、腾讯的混元大模型、百度的文心一言、月之暗面的Kimi、字节跳动的豆包等。

- **模型参数规模呈现指数级增长，模型性能持续提升。**近年来新推出的大语言模型所使用的数据量和参数规模呈现指数级增长，例如GPT-3模型参数约为**1750亿**，据Semianalysis推测GPT-4参数量达**1.8万亿**；同时，国内目前公布的大模型参数规模也普遍在**百亿至千亿级别**。性能方面，据Data Learner数据，GPT-4o在MMLU测评中获得88.7分的高分，分数较GPT-3大幅提高；国产模型中阿里的Qwen2.5-72B取得86.1分的高分，在各大模型中亦取得排名相对靠前的位置。

图：全球模型数量激增



资料来源：斯坦福大学《人工智能指数报告》，国信证券经济研究所整理

表：主流大模型信息对比

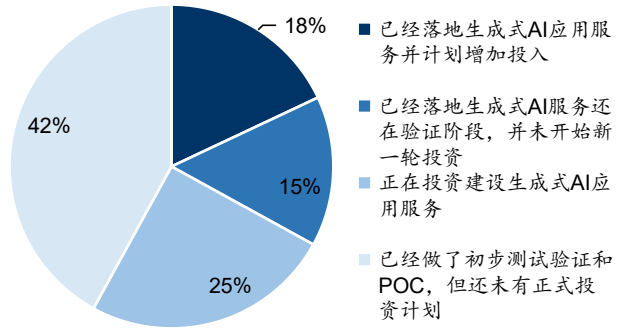
模型名称	参数大小（亿）	MMLU分数	发布者	发布时间	开源情况
GPT-4o	未公布	88.7	OpenAI	2024.5.13	未开源
Claude 3.5 Sonnet	未公布	88.7	Anthropic	2024.6.21	未开源
Claude 3-Opus	未公布	86.8	Anthropic	2024.3.4	未开源
GPT-4	未公布	86.4	OpenAI	2023.3.14	未开源
Qwen2.5-72B	727	86.1	阿里	2024.9.18	开源
Llama3.1-405B	4050	85.2	Meta	2024.7.23	开源
Gemini-Ultra	未公布	83.7	谷歌	2023.12.7	未开源
Qwen2.5-32B	320	83.3	阿里	2024.9.18	开源
Gemini 1.5 Pro	未公布	81.9	谷歌	2024.2.15	未开源
GLM4	未公布	81.5	智谱AI	2024.1.16	未开源
Grok-1.5	未公布	81.3	xAI	2024.3.29	未开源
YAYI2-30B	300	80.5	中科闻歌	2023.12.22	收费开源
Qwen1.5-110B	1100	80.4	阿里	2024.4.25	开源
Qwen2.5-14B	140	79.7	阿里	2024.9.18	开源
Llama3-70B	700	79.5	Meta	2024.4.18	开源
Gemini-Pro	1000	79.1	谷歌	2023.12.7	未开源
Claude 3-Sonnet	未公布	79.0	Anthropic	2024.3.4	未开源
DeepSeek-V2-236B	2360	78.5	DeepSeek	2024.5.6	开源
Qwen-72B	720	77.4	阿里	2023.11.30	开源
Yi-1.5-34B	340	77.1	零一万物	2024.5.13	开源
GPT-3.5	1750	70.0	OpenAI	2022.11.30	未开源
GPT-3	1750	53.9	OpenAI	2020.5.28	未开源

资料来源：Data Learner，国信证券研究所整理 注：MMLU是一种针对大模型的语言理解能力的测评，用以评测大模型基本的知识覆盖范围和理解能力。

AI模型已从大语言模型进化为全方位多模态模型，开启AI应用新纪元

- 23年3月以来，OpenAI所发布的GPT-4已经具备了多模态理解和多类型内容生成的能力，使得AI真正具备了重塑人机交互模式、全方位赋能人类生活的可能性。
- 24年12月OpenAI连续进行新品发布，包括具备多模态推理能力的完整版o1模型，正式发布Sora视频模型，开放并升级写作和编程工具Canvas，将ChatGPT与Apple生态深度整合、Siri与Apple Intelligence智能协同，发布了ChatGPT能够进行视频聊天的语音和视觉功能等。
- 字节跳动自24年5月豆包大模型家族正式发布到12月短短7个月时间，发布了Doubao-pro、Seed-TTS、Seed-ASR、Seed-Music、SeedEdit、视频生成模型、视觉理解模型等多项重磅成果，在语言能力、多模态理解与生成、模型推理、代码生成等方面不断提升。

图：中国市场企业生成式AI应用进度



资料来源：IDC，国信证券经济研究所整理

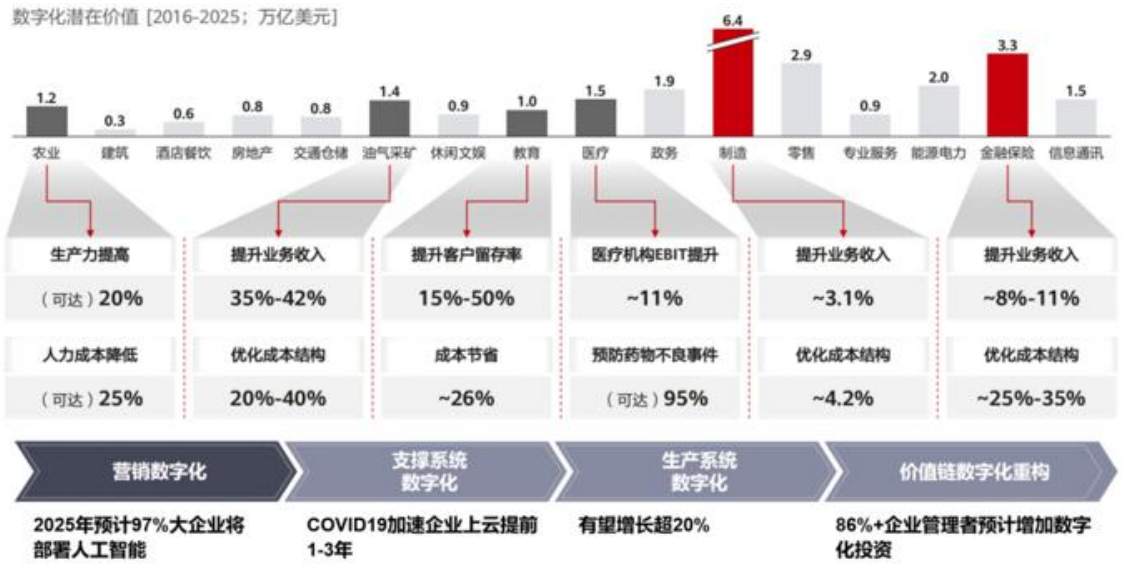
图：生成式AI市场概览

资料来源：IDC《市场概览：生成式AI技术和服务》，国信证券经济研究所整理

AI赋能下的“场景数字化”经济效益显著

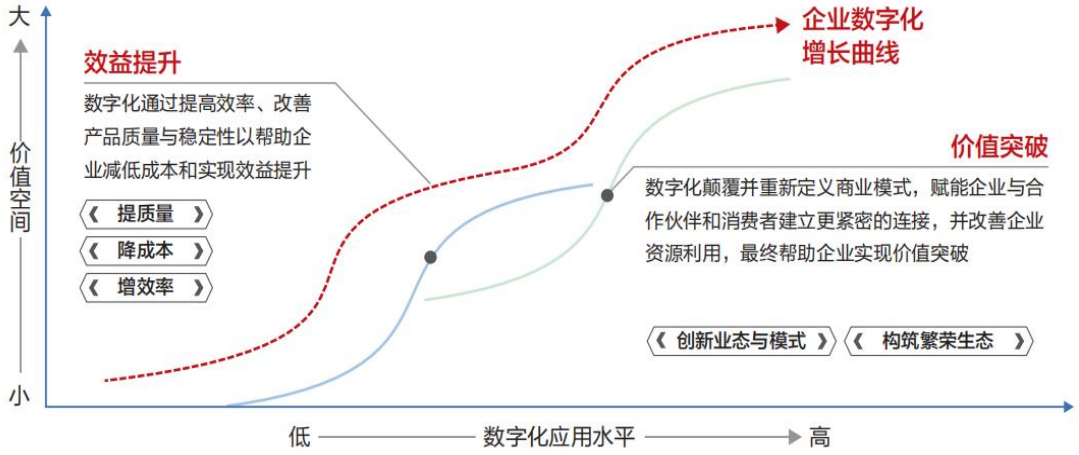
数字化解决方案的潜在价值对应近27万亿美元。数字化转型是以价值驱动的，其需求来源于企业即通过数字化来解决业务痛点、创造真实价值。根据华为的数据，制造业、金融保险、零售、能源电力等产业的数字化潜在价值均在2万亿美元以上；以作为支柱性工业的制造业为例，多为重资产企业，且流程复杂，需要在制造、运输、管理等多个环节进行数字化应用以实现降本增效，转型诉求强，数字化创造的潜在价值达6万亿美元。

图：数字化潜在价值



资料来源：华为《数字化转型，从战略到执行》，罗兰贝格，国信证券经济研究所整理

图：制造业企业数字化演进



资料来源：华为《加速行业智能化白皮书》、《数字化转型，从战略到执行》，国信证券经济研究所整理

AI赋能下的“场景数字化”经济效益显著

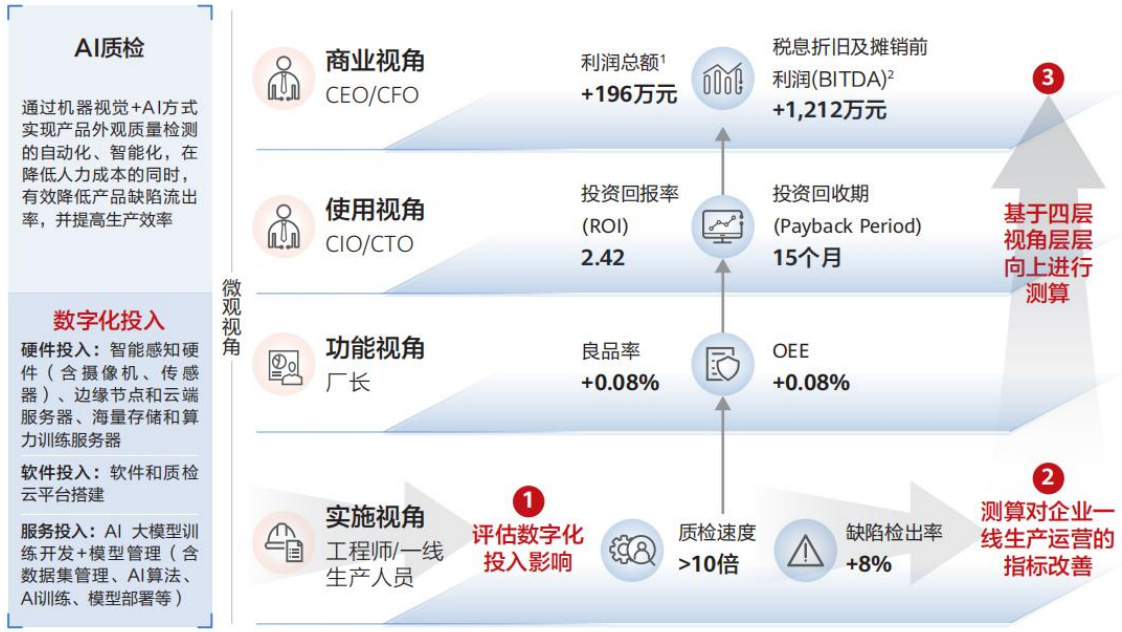
国信证券 GUOSEN SECURITIES

随着企业对数据的需求从收集到理解并进行应用过渡，AI是挖掘数据价值的重要工具。根据华为的数据，家电行业的大模型-AI质检系统借助AI能力，质检速度提升了数十倍。通过机器视觉＋AI的方式实现产品外观质量检测的自动化、智能化，系统能够使得检测速度提升10倍以上，缺陷检出率达到98%，进一步提高了质检质量，提高良品率，并且帮助企业节约因质量问题产生的退换货成本。同时，该系统能够帮助企业大幅减少质检工时，降低了人力成本。

图：制造业的数字化转变

资料来源：华为，国信证券经济研究所整理

图：家电AI质检应用案例



资料来源：华为，国信证券经济研究所整理

AI推动全球IT支出增长，生成式AI市场规模持续提高

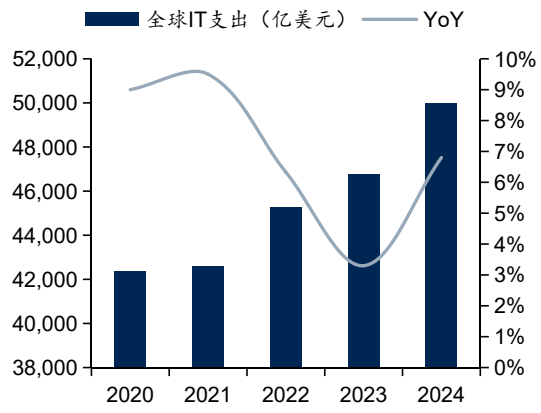
国信证券 GUOSEN SECURITIES

- **AI技术在企业端部署和应用推动全球IT支出的增长。**AI技术的发展驱动企业业务自动化、帮助企业优化资源配置并协助数据分析，从而提升业务流程效率，推动企业的智能化转型需求。个性化服务和智能应用带来了新的市场需求，以AIGC为代表的AI应用需要强大的计算能力和数据存储，推动了高性能计算、数据管理以及云计算的相关投资。随着AI应用的增多，企业对网络安全和合规服务的需求也在增加，推动相关领域的IT支出增长。据Gartner预测，企业机构将于2024年加快投资于使用生成式AI，2024年全球IT总支出预计将达到5万亿美元，较2023年增长6.8%；中国IT总支出预计将达到3.9万亿元，较2023年增长6.2%。

- **AI技术将产生巨大的经济影响，其投入产出效益显著。**据IDC数据，预计到2030年，人工智能对全球经济的累计影响将达到19.9万亿美元，占到预计2030年全球GDP的3.5%。到2030年，每在AI解决方案和服务上花费1美元，将产生4.6美元的经济效益，包括直接影响和间接影响。

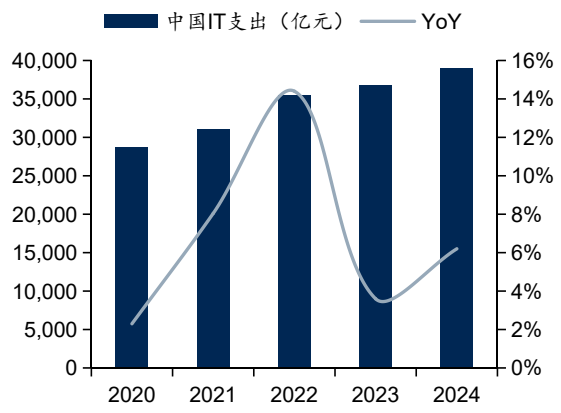
- **生成式AI市场将成为当前最热门的IT领域。**据IDC数据，24年中国生成式AI市场预计将达到33亿美元，预计到2028年将达到135亿美元，2024-2028年复合增长率将达33.4%，同时生成式AI市场规模占到整体AI市场规模的比例将由16%上升到29%。对于企业来讲，对于生成式AI的支出亦将经历不同阶段的重点，例如2024-2025年，支出主要集中在生成式AI基础设施建设；2025-2026年，支出重点用于推进生成式AI平台与解决方案建设；2027年及以后，支出重点着力于生成式AI服务。

图：2020-2024年全球IT支出情况



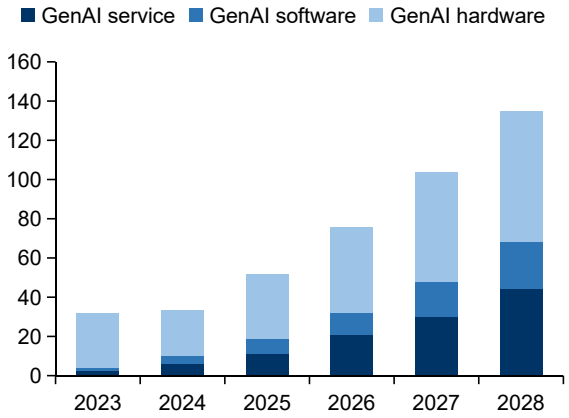
资料来源：Gartner，国信证券经济研究所整理

图：2020-2024年中国IT支出情况



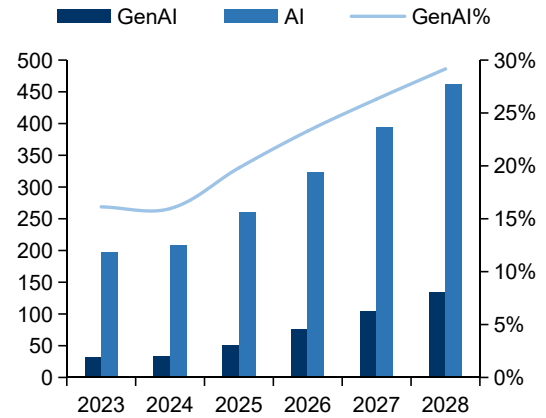
资料来源：Gartner，国信证券经济研究所整理

图：2023-2028年中国生成式AI市场预测（亿美元）



资料来源：IDC's Worldwide AI and Generative AI Spending Guide V2，2024，国信证券经济研究所整理

图：2023-2028年中国生成式AI与整体AI市场规模预测（亿美元）



资料来源：IDC's Worldwide AI and Generative AI Spending Guide V2，2024，国信证券经济研究所整理

智能算力是构建大模型的重要底座，AI算力需求持续攀升

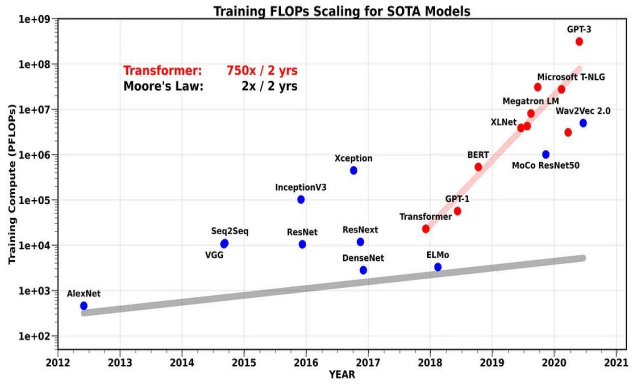
国信证券 GUOSEN SECURITIES

- **大模型训练、AI应用需求兴起，推动全球算力需求快速增长。**全球算力需求飙升主要基于以下原因：1）模型能力提升依赖更大的训练数据量和参数量，对应更高的算力需求；2）AI模型的发展方向转向多模态，训练模型的数据从单一文字数据发展到目前的图片、视频数据，均需要更强的算力处理；3）模型种类多样化（文生图、文生视频）以及新推出的模型数量激增，均推动算力需求的增长，以AIGC为代表的AI应用用户数量爆发，推理侧算力需求快速增长。

- **全球数据总量大幅上涨，数据中心算力需求快速增长。**随着人工智能等新技术发展，海量数据的产生及其计算处理成为数据中心发展关键。据IDC数据，全球数据总量预计由2021年的82.47 ZB上升至2026年的215.99 ZB，对应CAGR达21.24%。其中，大规模张量运算、矩阵运算是人工智能在计算层面的突出需求，高并行度的深度学习算法在视觉、语音和自然语言处理等领域上的广泛应用使得算力需求呈现指数级增长。此外，据IDC数据，中国生成式AI日均Tokens处理规模显著增长，预计中国生成式AI日均Tokens调用量到2024年底将达到每天1.12万亿，是2023年底每天35亿规模的320倍。

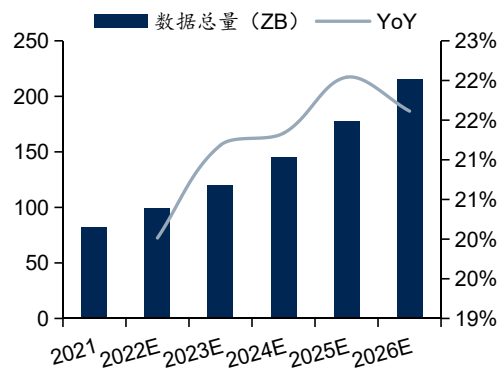
- **智能算力是构建大模型的重要底座，以AI服务器为代表的全球智能算力需求激增。**算力可分为通用算力、智能算力及超算算力：1）通用算力：由基于CPU的服务器提供算力，主要用于基础通用计算；2）智能算力：由基于GPU、FPGA、ASIC等AI芯片的加速计算平台提供的算力，主要用于人工智能训练和推理计算；3）超算算力：由超级计算机等高性能计算集群提供算力，主要用于尖端科学领域的计算。早期通用算力占整体算力的比重达90%以上，随着人工智能技术的发展，智能算力规模迅速增长。**据中国信息通信研究院预期，2030年全球智能算力规模将达52.5ZFLOPS。据IDC预期，2023年中国智能算力规模达414.1EFLOPS，至2027年将达1117.4EFLOPS。**

图：AI大模型所需算力平均每2年增长750倍



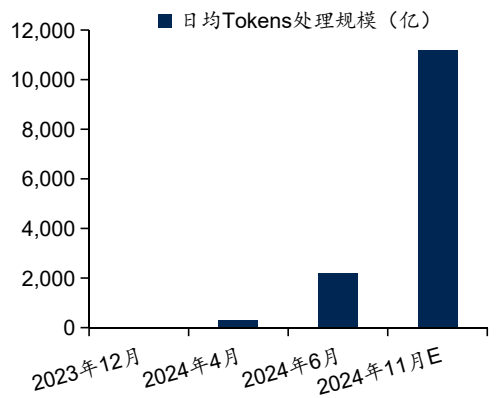
资料来源：riselab，国信证券经济研究所整理

图：2021-2026年全球数据总量及预测



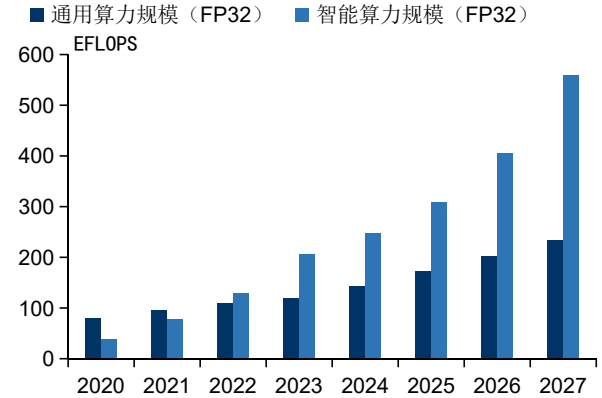
资料来源：IDC，国信证券经济研究所整理

图：中国生成式AI日均Tokens处理规模



资料来源：IDC，国信证券经济研究所整理

图：中国算力规模及预期（单位：EFLOPS）



资料来源：IDC，国信证券经济研究所整理

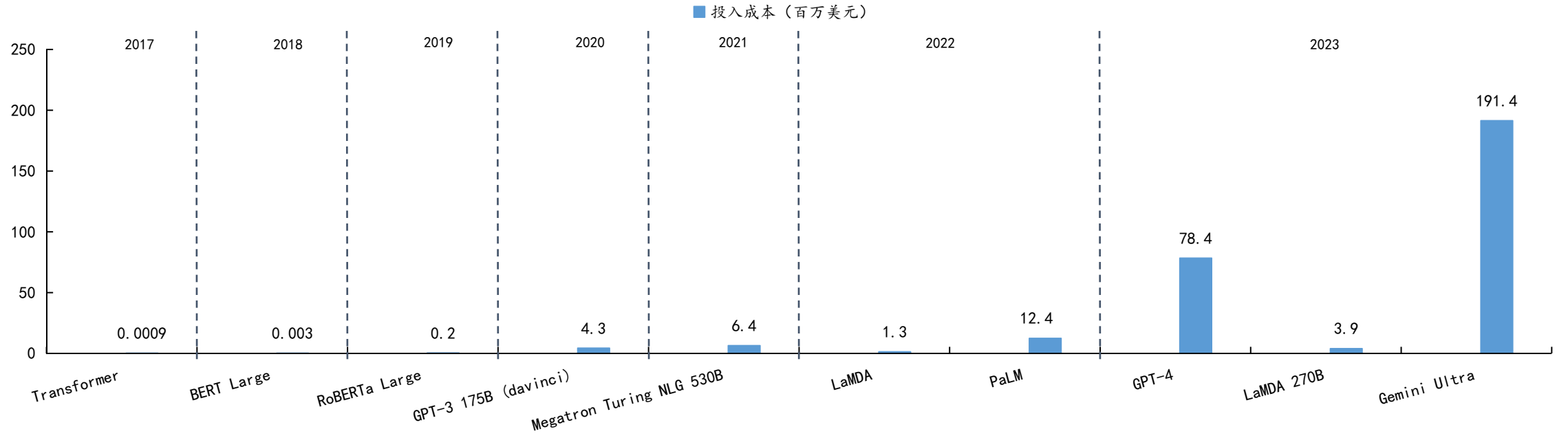
AI应用智能化推动算力基础设施升级，算力厂商将率先受益

国信证券 GUOSEN SECURITIES

- **AI应用智能化推动算力基础设施升级。**伴随着AI应用的智能化，一方面将通过优化智能汽车、智能机器人、智能家居、空间计算终端（MR\VR\AR）等各类智能物联产品的人机交互体验，加速其市场推广速度；另一方面也将倒逼相应的算力基础设施、终端硬件架构为此做出适应性的升级。

- **算力需求催化投资，算力厂商将率先受益。**根据斯坦福大学《人工智能指数报告》估算，OpenAI的GPT-4使用了价值约7800万美元的计算资源进行训练，而谷歌的Gemini Ultra耗费了1.9亿美元的计算成本。2024年3月，微软和OpenAI宣布计划投资1000亿美元打造星际之门AI超算，全球算力投资迅速提升，算力厂商将率先受益。

图：全球模型训练投入激增



资料来源：斯坦福大学《人工智能指数报告》，国信证券经济研究所整理

英伟达CUDA平台及GPU架构快速迭代更新奠定其领先地位

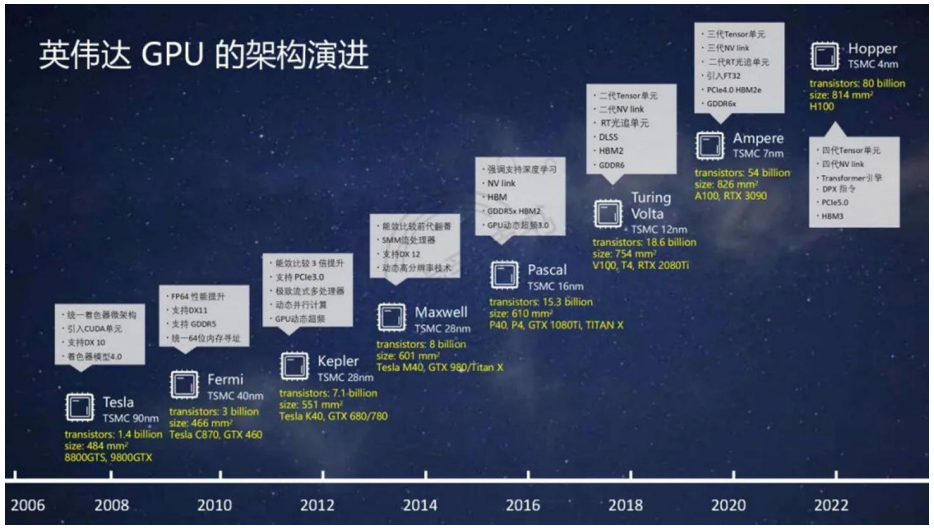
国信证券 GUOSEN SECURITIES

- **英伟达将GPU从图形处理器演进为通用计算处理器，CUDA降低通用GPU应用门槛。**成立于1993年的英伟达以制造单芯片图形用户界面加速器起家，于1999年发明了图形处理器GPU，从而定义了现代计算机图形学，并确立在该领域的领导地位。2006年公司推出用于通用GPU计算的CUDA平台，**是首次可以利用GPU作为C语言编译器的开发环境，**使得GPU能够进行图像处理之外的通用计算，英伟达GPU体系结构全面支持通用编程，GPU成为了真正的GPGPU（通用GPU）。

- **英伟达GPU加速计算发展始于Tesla架构，其架构约每两年完成迭代更新奠定其领先地位。**

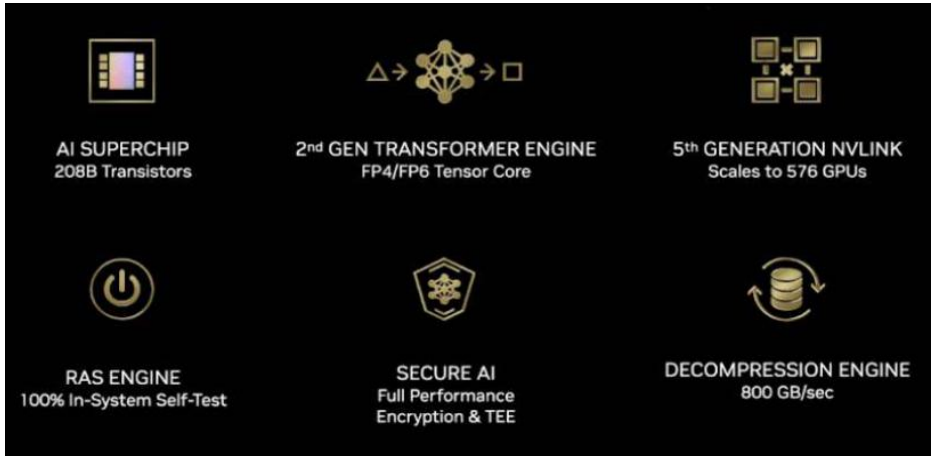
- 2008年，Tesla架构推出，成为第一代真正开始用于并行运算的GPU架构。
- 2010年，Fermi架构推出，是第一个支持DirectX 11的GPU计算架构，采用台积电40nm制程。
- 2012年的Kepler架构是Fermi的升级版，整体架构保持一致性，采用台积电28nm制程。
- 2014年的Maxwell架构通过优化架构，提供了可观的能耗比提升。
- 2016年，Pascal架构推出，采用台积电16nm制程，支持DirectX 12标准，是首个为深度学习而设计的GPU架构。
- 2017年的Volta架构专注于提高深度学习的性能，采用台积电12nm制程，
- 2018年的Turing架构是全球首款支持实时光线追踪的GPU架构。
- 2020年，Ampere架构推出，采用台积电7nm/三星8nm制程，统一了AI训练和推理，并在光线追踪和DLSS（深度学习超级采样）方面有显著的改进。
- 2022年的Hopper架构采用台积电4nm制程，集成多达800亿个晶体管，主要面向AI及数据中心等构建。
- 2024年3月最新推出的Blackwell架构采用台积电4nm制程，集成了2080亿个晶体管，使用了二代Transformer、Secure AI、5代NVLink等最新技术。

图：英伟达GPU架构演进历程



资料来源：woshipm，国信证券经济研究所整理

图：英伟达Blackwell架构的技术突破



资料来源：英伟达官网，国信证券经济研究所整理

随着芯片架构不断演进，英伟达GPU算力成倍增长

国信证券 GUOSEN SECURITIES

- **英伟达GPU芯片随着架构的不断演进及算力的成倍增长，在大算力需求的AI大模型训练中得到广泛运用。**基于Ampere架构的A100 GPU建立在Volta和Turing SM架构中引入的特性之上，并显著提高了性能，与Volta和Turing相比，每平方米的计算马力增加了2倍；Ampere架构还引入了细粒度结构稀疏性，可以使深层神经网络的计算吞吐量翻倍。Hopper架构利用专为加速AI模型训练而设计的Transformer引擎，进一步提升Tensor核心技术。Hopper Tensor核心可应用混合式FP8和FP16精确度，大幅加速Transformer的AI运算；和前一代Ampere相比，Hopper将TF32、FP64、FP16和INT8每秒浮点运算次数提高三倍。Blackwell架构使用了第二代Transformer引擎，将定制的Blackwell Tensor Core技术与NVIDIA TensorRT-LLM和NeMo框架创新相结合，加速大语言模型和专家混合模型的推理和训练；与上一代H100相比，使用Blackwell架构的GB200 NVL72将资源密集型应用程序（例如1.8T参数GPT-MoE）的速度提高了30倍。

表：英伟达数据运算GPU主流产品性能

	B200	B100	H200 SXM	H100 SXM	H800 SXM	A100 SXM	A800 SXM	L40S	L40
FP4	18 PFLOPS	14 PFLOPS	-	-	-	-	-	-	-
INT4	-	-	-	-	-	-	-	1466 TOPS	1448 TOPS
FP8/FP6	9 PFLOPS	7 PFLOPS	3958 TFLOPS	3958 TFLOPS	3958 TFLOPS	-	-	1466 TFLOPS	724 TFLOPS
INT8	9 POPS	7 POPS	3958 TOPS	3958 TOPS	3958 TOPS	1248 TOPS	1248 TOPS	1466 TOPS	724 TFLOPS
FP16	4.5 PFLOPS	3.5 PFLOPS	1979 TFLOPS	1979 TFLOPS	1979 TFLOPS	624 TFLOPS	624 TFLOPS	733 TFLOPS	362.1 TFLOPS
TF32	2.2 PFLOPS	1.8 PFLOPS	989 TFLOPS	989 TFLOPS	989 TFLOPS	312 TFLOPS	312 TFLOPS	366 TFLOPS	191 TFLOPS
FP32	80 TFLOPS	60 TFLOPS	67 TFLOPS	67 TFLOPS	67 TFLOPS	19.5 TFLOPS	19.5 TFLOPS	91.6 TFLOPS	90.5 TFLOPS
FP64	40 TFLOPS	30 TFLOPS	34 TFLOPS	34 TFLOPS	1 TFLOPS	9.7 TFLOPS	9.7 TFLOPS	-	-
显存	最高192GB	最高192GB	141GB	80GB	80GB	80GB	80GB	48GB	48GB
显存带宽	最高8 TB/s	最高8 TB/s	4.8 TB/s	3.35 TB/s	3.35 TB/s	2039 GB/s	2039 GB/s	864 GB/s	864 GB/s
热设计功耗	1000W	700W	最高700W	最高700W	最高700W	400W	400W	350W	300W
互联速度	NVLink：1.8TB/s PCIe 6.0：256GB/s	NVLink：1.8TB/s PCIe 6.0：256GB/s	NVLink：900GB/s PCIe 5.0：128GB/s	NVLink：900GB/s PCIe 5.0：128GB/s	NVLink：400GB/s PCIe 5.0：128GB/s	NVLink：600GB/s PCIe 4.0：64GB/s	NVLink：400GB/s PCIe 4.0：64GB/s	PCIe 4.0：64GB/s	PCIe 4.0：64GB/s

资料来源：英伟达官网，国信证券经济研究所整理

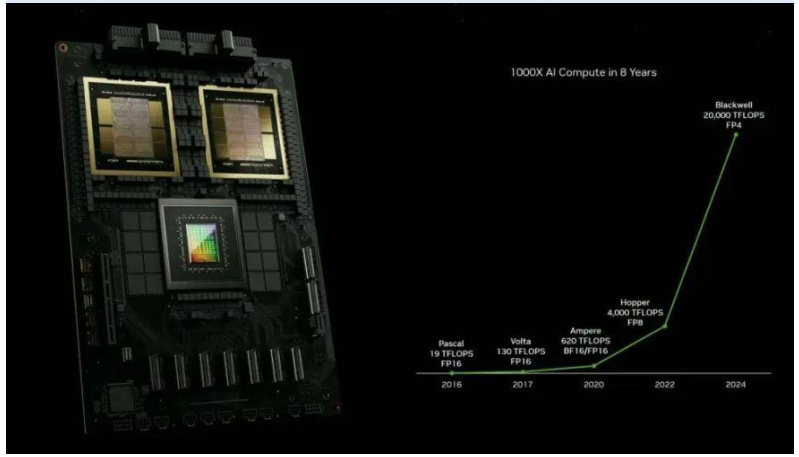
GB200系统强势赋能下一代AI模型，系列新品即将陆续出货

国信证券 GUOSEN SECURITIES

- **面向生成式AI时代的全新机架级扩展的DGX SuperPOD架构基于DGX GB200系统，将前所未有赋能下一代AI模型。**GB200是由两个Blackwell B200 GPU和一个Grace CPU组成的AI加速平台，每个B200 GPU含有2080亿个晶体管。相较于H100，GB200的算力提升了6倍；而在处理多模态特定领域任务时，其算力更是达到H100的30倍。GB200 NVL72是一套多节点液冷机架级扩展系统，适用于高度计算密集型的工作负载，它将36个Grace Blackwell超级芯片组合在一起，其中包含通过第五代NVLink相互连接的72个Blackwell GPU和36个Grace CPU。DGX SuperPOD由8个或以上的DGX GB200 NVL72系统构建而成，这些系统通过NVIDIA Quantum InfiniBand网络连接，可扩展到数万个GB200超级芯片，可以用于处理万亿参数模型，能够保证超大规模生成式AI训练和推理工作负载的持续运行。

- **预计B200和GB200系列在2024年第四季度和2025年第一季度之间陆续出货，B300系列将于2025年第二季度至第三季度之间陆续出货。**据TrendForce数据，英伟达对Blackwell系列芯片的划分更为细致，以向大型云服务商提供符合其能效要求和服务器OEM性价比需求的产品，并根据供应链情况动态调整。预计2025年英伟达将更着力于营收贡献度较高的AI机种，例如积极投入技术和资源在NVL Rack方案，协助服务器系统厂商针对NVL72系统调教或液冷散热等，推动大型云服务厂商从现有NVL36转为扩大导入NVL72。出货占比方面，据TrendForce数据，英伟达高端GPU增长明显，预计2024年出货占比约为50%；预计2025年受Blackwell新平台带动，其高端GPU出货占比将提升至65%以上。TrendForce指出，英伟达近期将其所有Blacwell Ultra产品更名为B300系列，预计B200和GB200系列在2024年第四季度和2025年第一季度之间陆续出货，B300系列将于2025年第二季度至第三季度之间陆续出货。

表：英伟达GB200芯片及性能提升示意图



资料来源：英伟达官网，国信证券经济研究所整理

表：英伟达DGX SuperPOD架构示意图

资料来源：英伟达官网，国信证券经济研究所整理

表：英伟达Blackwell系列产品重要规格预测

旧名称	新名称	主要服务器出货单位	HBM类型	CoWoS类型
B100	B100	HGX	HBM3e 8hi*8(192GB)	CoWoS-L
B200	B200	HGX	HBM3e 8hi*8(192GB)	CoWoS-L
B200 Ultra	B300	HGX	HBM3e 12hi*8(288GB)	CoWoS-L
GB200	GB200	NVL72(main)、NVL36	HBM3e 8hi*8(192GB)	CoWoS-L
GB200 Ultra	GB300	NVL72(main)、NVL36	HBM3e 12hi*8(288GB)	CoWoS-L
B200A Ultra	B300A	HGX、MGX	HBM3e 12hi*4(144GB)	CoWoS-S
GB200A Ultra	GB300A	NVL36、MGX	HBM3e 12hi*4(144GB)	CoWoS-S

资料来源：TrendForce，国信证券经济研究所整理

国内外云服务商资本开支快速增长，国内大厂增速明显

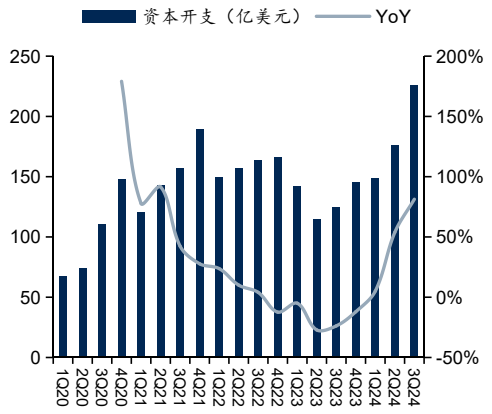
国信证券 GUOSEN SECURITIES

- **国内外大型云服务厂商近两年资本开支快速增长，算力“军备竞赛”愈演愈烈。**国外四大CSP厂商今年前三季度资本开支均已超过200亿美元，亚马逊更是超过500亿美元。中国头部云服务商如腾讯、阿里巴巴等今年前三季度资本开支增长均超过100%。

- 国外四大CSP厂商亚马逊、微软、谷歌、Meta在2024年第三季度资本开支分别达到226.2亿、149.23亿、130.61亿、82.58亿美元，同比分别增长81.3%、50.5%、62.1%、26.2%；2024年前三季度累计资本开支分别达551.65亿、397.48亿、382.59亿、228.31亿美元，同比分别增长44.6%、56.1%、80.2%、16.5%。

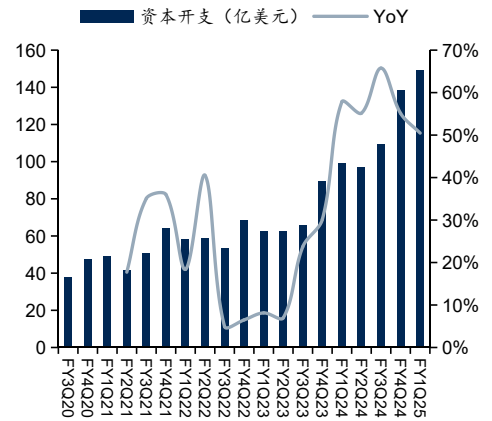
- 国内头部云服务商如腾讯、阿里巴巴在2024年第三季度资本开支分别达到170.94亿、169.77亿元，同比分别增长113.54%、312.86%；2024年前三季度累计资本开支分别达到401.82亿、390.90亿元，同比分别增长145.5%、209.5%。

图：亚马逊季度资本开支



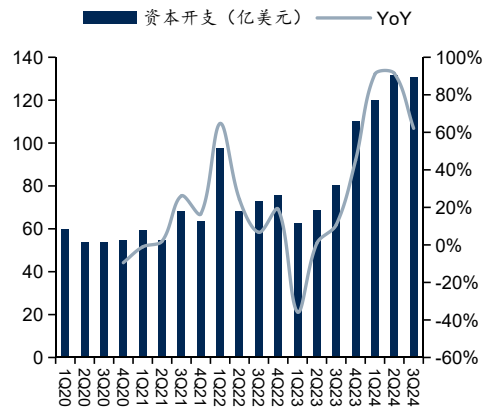
资料来源：Wind，国信证券经济研究所整理

图：微软季度资本开支



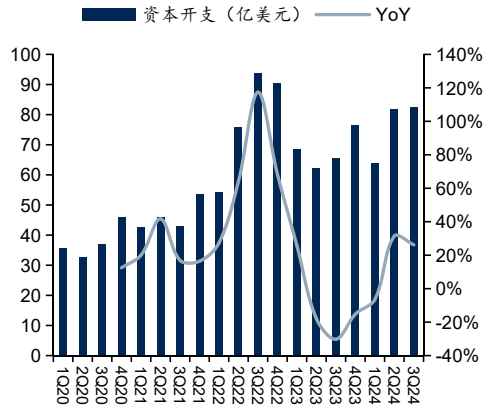
资料来源：Wind，国信证券经济研究所整理

图：谷歌季度资本开支



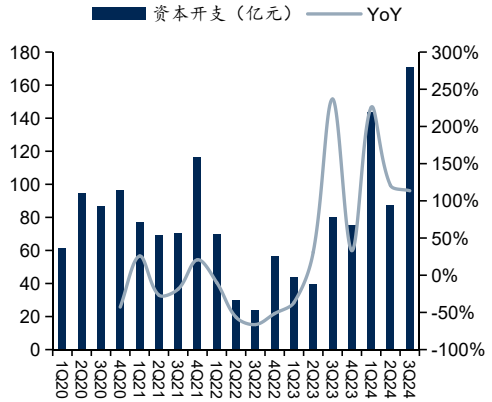
资料来源：Wind，国信证券经济研究所整理

图：Meta季度资本开支



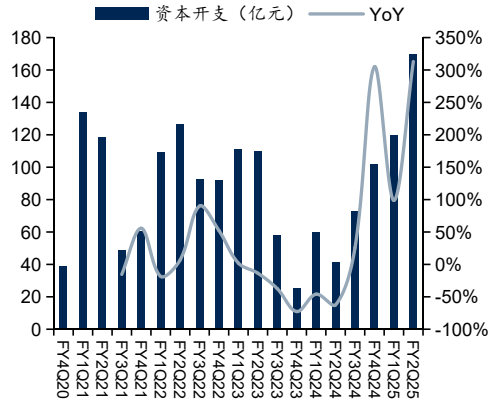
资料来源：Wind，国信证券经济研究所整理

图：腾讯季度资本开支



资料来源：Wind，国信证券经济研究所整理

图：阿里巴巴季度资本开支



资料来源：Wind，国信证券经济研究所整理

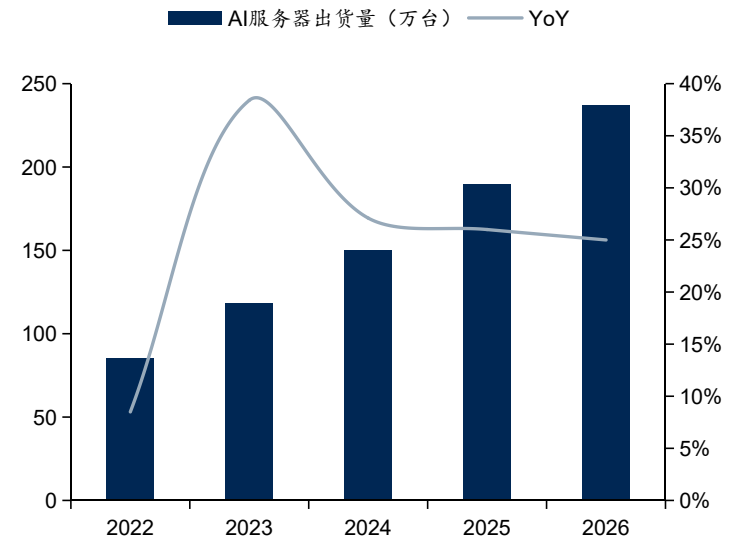
算力需求爆发式增长，AI服务器市场规模大幅提升

国信证券 GUOSEN SECURITIES

- **受益于智能算力市场的推动，全球AI服务器市场规模实现快速增长。**据TrendForce数据，预计2024年全球AI服务器市场规模为1870亿美金，同比增长69%；从服务器出货量占比来看，预计2024年AI服务器占比为12.2%，同比提升3.4pct。TrendForce预计AI服务器出货量将由2023年的118万台增长至2026年的237万台，对应CAGR为26%。假设单台AI服务器价值量为25万美金，则预计2026年AI服务器市场规模为5922.5亿美金。

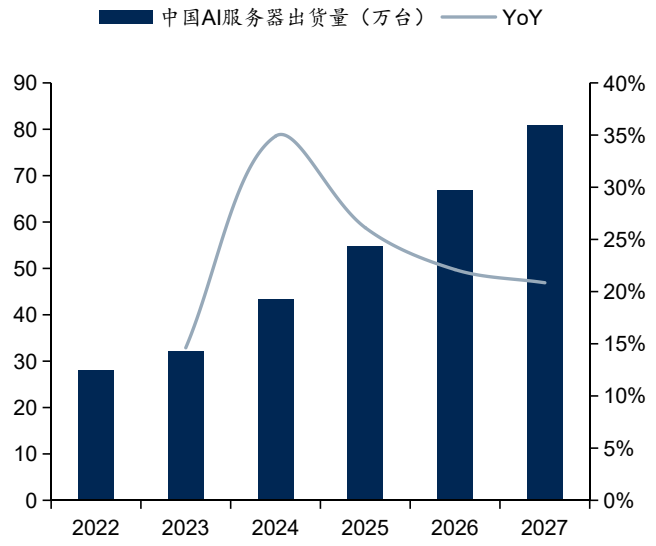
- **中国AI服务器市场规模同样将实现快速增长，AI服务器工作负载将由训练逐步过渡到推理。**据IDC数据，2023年中国AI服务器出货量达32.2万台，预计到2027年将达到80.9万台，对应CAGR达25.9%；对应到2023年AI服务器市场规模为60.8亿美元，预计到2027年将达到134亿美元，对应CAGR达21.8%。从工作负载来看，2023年训练服务器占比达58.7%。随着训练模型的完善与成熟，模型和应用产品逐步进入投产模式，处理推理工作负载的人工智能服务器占比将随之攀升，到2027年，用于推理的工作负载将达到72.6%。

图：全球AI服务器出货量及预测



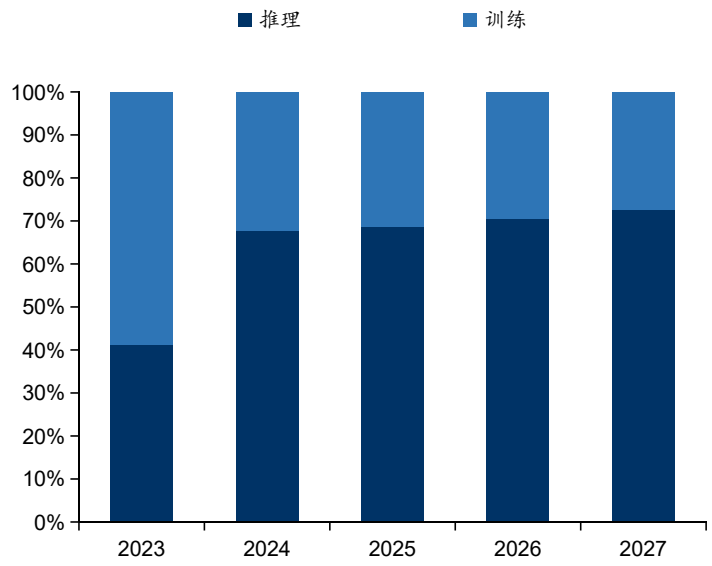
资料来源：Trendforce，国信证券经济研究所整理

图：中国AI服务器出货量及预测



资料来源：IDC，国信证券经济研究所整理

图：中国AI服务器工作负载占比及预测



资料来源：IDC，国信证券经济研究所整理

政策支持亦将拉动中国AI服务器市场规模增长

国信证券 GUOSEN SECURITIES

● **政策支持亦将拉动中国AI服务器市场规模增长。**在当前数字经济时代背景下，国家出台多个政策支持AI产业发展，AI服务器行业将保持快速增长。相关企业加速布局以及人工智能应用场景的逐步落地，AI服务器在服务器整体市场中比重提高。中国的企业和研究机构积极进行人工智能服务器的技术研发和创新，包括高性能处理器、大容量内存、高速存储器和高效冷却系统等领域的创新，以满足计算能力和数据处理速度的需求。

表：中国人工智能行业政策节选

发布日期	发布单位	政策名称	主要内容
2024年1月	工信部	《国家人工智能产业综合标准化体系建设指南》（征求意见稿）	到2026年，共性关键技术和应用开发类计划项目形成标准成果的比例达到60%以上，标准与产业科技创新的联动水平持续提升。新制定国家标准和行业标准50项以上，推动人工智能产业高质量发展的标准体系加快形成。开展标准宣贯和实施推广的企业超过1000家，标准服务企业创新发展的成效更加凸显。参与制定国际标准20项以上，促进人工智能产业全球化发展。
2023年4月	工信部、中央网信办、国家发改委教育部等	《关于推进IPv6技术演进和应用创新发展的实施意见》	推动IPv6与5G、人工智能、云计算等技术的融合创新，支持企业加快应用感知网络、新型IPv6测量等“IPv6+”创新技术在各类网络环境和业务场景中的应用。
2023年2月	中共中央、国务院办公厅	《质量强国建设纲要》	加快大数据、网络、人工智能等新技术的深度应用，促进现代服务业与先进制造业、现代农业融合发展。
2022年12月	中共中央、国务院办公厅	《扩大内需战略规纲要(2022-2035年)》	加快物联网、工业互联网、卫星互联网、壬兆光网建设，构建全国一体化大数据中心体系，布局建设大数据中心国家枢纽节点，推动人工智能、云计算等广泛、深度应用，促进“云、网、端”资源要素相互融合、智能配置。推动5G、人工智能、大数据等技术。
2022年8月	科技部	《关于支持建设新一代人工智能示范应用场景的通知》	充分发挥人工智能赋能经济社会发展的作用，围绕构建全链条、全过程的人工智能行业应用生态，支持一批基础较好的人工智能应用场景，加强研发上下游配合与新技术集成，打造形成一批可复制、可推广的标杆型示范应用场景。首批支持建设十个示范应用场景。
2022年7月	科技部、教育部、工业和信息化部、交通运输部等	《关于加快场景创新以人工智能高水平应用促进经济高》	场景创新成为人工智能技术升级、产业增长的新路径，场景创新成果持续涌现推动新一代人工智能发展上水平。鼓励在制造、农业、物流、金融、商务、家等重点行业深入挖据人工智能技术应用场景，促进智能经济高端高效发展。
2021年5月	国家发改委、中央网信办、工信部中央能源局	《全国一体化大数据中心协同创新体系算力枢纽实施方案》	引导超大型、大型数据中心集聚发展，构建数据中心集群，推进大规模数据的“云端”分析处理，重点支持对海量规模数据的集中处理，支撑工业互联网、金融证券、灾害预远程医疗、视频通话、人工智能推理等抵近一线、高频实时交互型的业务需求，数据中心端到端单向网络时延原则上在20毫秒范围内。
2021年3月	中共中央	《国民经济和社会发展第十四个五年规划和二零三五年远景目标》	瞄准人工智能等前沿领域，实施一批具有前瞻性、战略性的国家重大科技项目。推动互联网、大数据、人工智能等同各产业深度融合，推动先进制造业集群发展，构建一批各具特色、优势互补、结构合理的战略性新兴产业增长引擎，培育新技术、新产品、新业态、新模式。

资料来源：中商产业研究院，国信证券经济研究所整理

AI服务器搭载AI芯片仍以GPU为主，英伟达占据绝对的供应地位

国信证券 GUOSEN SECURITIES

- **AI服务器搭载AI芯片仍以GPU为主，搭载ASIC芯片服务器占比有上升趋势。**当前主流的AI芯片包括GPU、FPGA、ASIC等，其中GPU是前期较为成熟的芯片架构，属于通用型芯片；ASIC属于为AI特定场景定制的芯片。由于GPU通用型较强、适合大规模并行运算，设计和制造工艺成熟，适用于高级复杂算法和通用性人工智能平台。由于ASIC根据产品的需求进行特定设计和制造的集成电路，能够更有针对性地进行硬件层次的优化，因此具有更高的处理速度和更低的能耗；相比于其他AI芯片，ASIC设计和制造需要大量的资金、较长的研发周期和工程周期。据TrendForce数据，预计2024年搭载GPU的AI服务器占比约为71%，仍占据主导地位。而随着北美云服务商如亚马逊、Meta等，以及国内云服务商如阿里、百度、华为等持续积极扩大自研ASIC方案，使得搭载ASIC服务器占整体AI服务器比重在2024年将提升至26%。

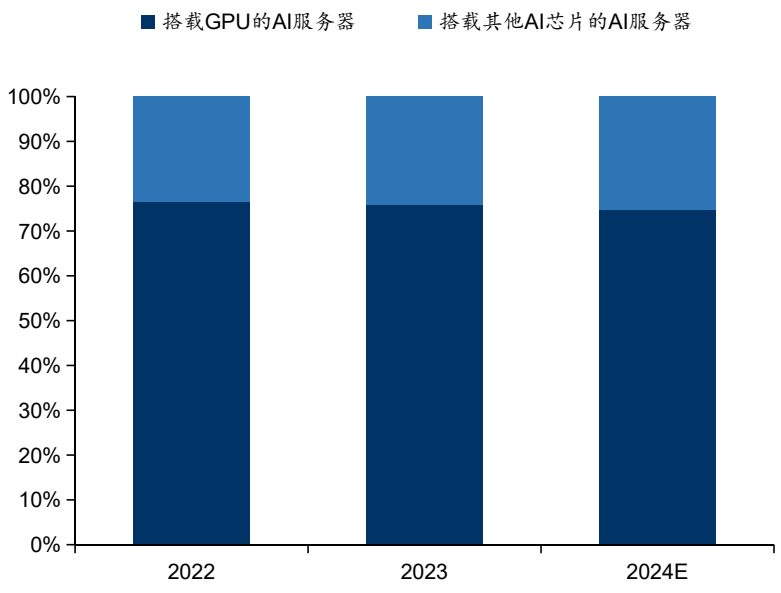
- **英伟达仍是搭载GPU的AI服务器的绝对芯片供应商。**据TrendForce数据，单看AI服务器搭载GPU的芯片供应商中，英伟达占据绝对的主导地位，2022-2024年市占率均达到85%以上。随着AMD发布Instinct系列AI芯片并在AI服务器方面不断发力，其市占率有望从2022年的5.7%上升至2024年的8.1%。Intel在AI服务器芯片供应商中占比近年保持相对稳定，约占3%左右。

表：不同技术架构AI芯片比较

AI芯片种类	GPU	ASIC
定制化程度	通用型	全定制化
算力	中	高
价格	高	低
优点	通用型较强、适合大规模并行运算；设计和制造工艺成熟。	通过算法固化实现极致的性能和能效、平均性强；功耗低；体积小；量产后成本低。
缺点	并行运算能力在推理段无法完全发挥。	前期投入成本高；研发时间长；技术风险大。
应用场景	高级复杂算法和通用性人工智能平台。	当客户处在某个特殊场景，可以为其独立设计一套专业智能算法软件。

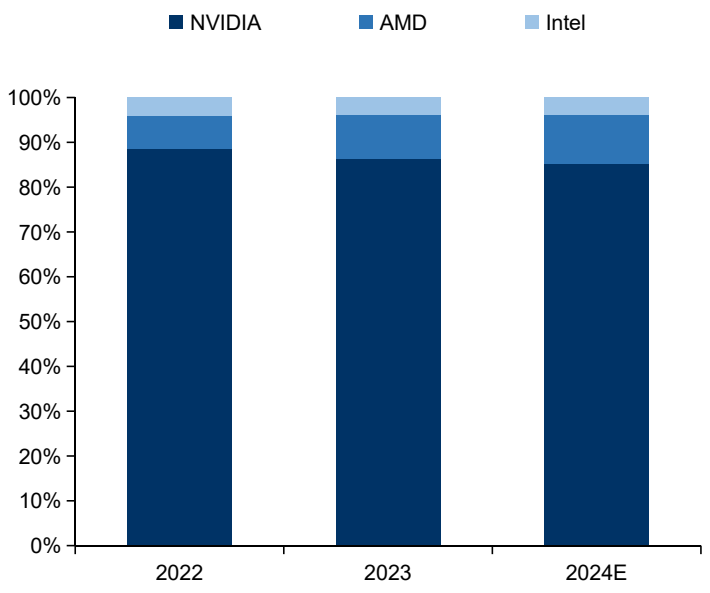
资料来源：亿欧智库，国信证券经济研究所整理

图：搭载不同AI芯片的AI服务器占比



资料来源：Trendforce，国信证券经济研究所整理

图：搭载GPU的AI服务器市场格局



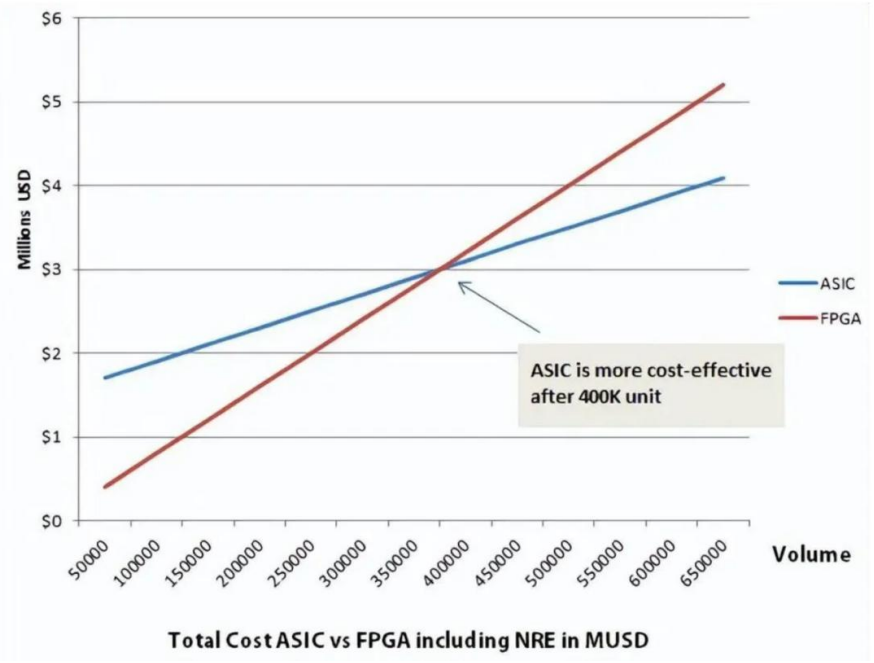
资料来源：Trendforce，国信证券经济研究所整理

牧本定律摆向定制化，关注国产ASIC服务商

国信证券 GUOSEN SECURITIES

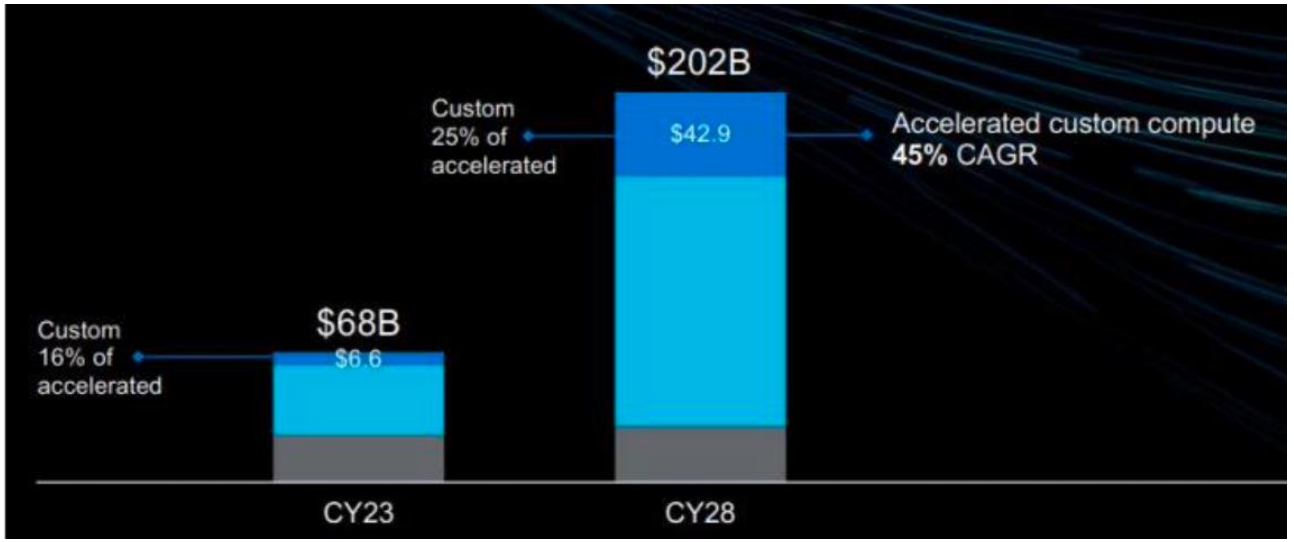
- ASIC专用集成电路是应特定用户的要求，或特定电子系统的需要，专门设计、制造的集成电路。根据下图显示，40万片的产量是ASIC和FPGA成本高低的分界线，当产量大于40万片时，ASIC的性价比相对FPGA更高。
- 根据Marvell预测，数据中心定制加速芯片2023至2028年市场规模CAGR有望达到45.5%。2023年数据中心ASIC市场规模约66亿美元，占整体数据中心加速计算芯片680亿美元市场的16%。预计到2028年数据中心ASIC市场将达到429亿美元，占整体数据中心加速芯片2020亿美元的25%。相较于GPU，AI ASIC整体复合增速更快，达到45.4%。

图：ASIC在达到40万片后性价比相对FPGA更高



资料来源：鲜枣课堂公众号，国信证券经济研究所整理

图：数据中心定制加速计算芯片市场规模



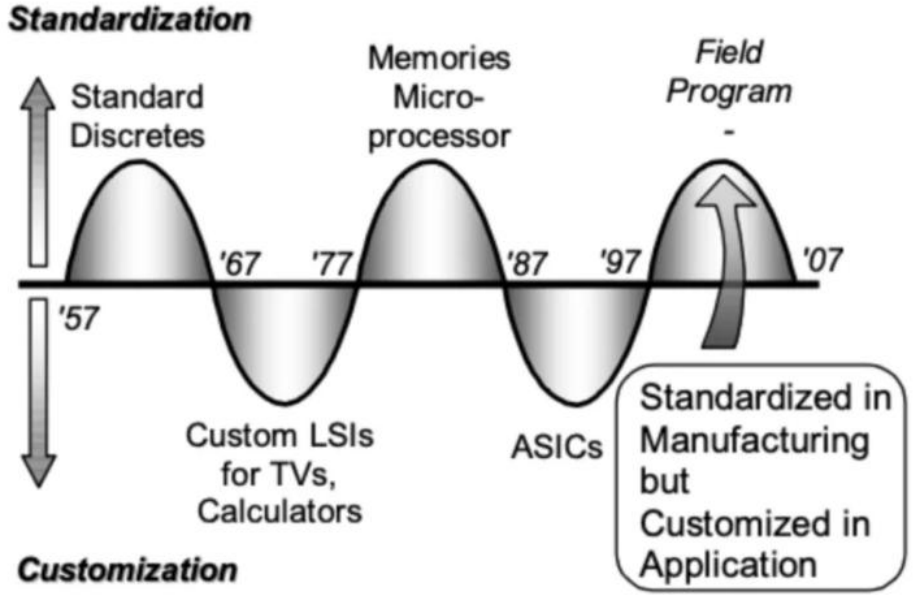
资料来源：Marvell，国信证券经济研究所整理

牧本定律摆向定制化，关注国产ASIC服务商

国信证券 GUOSEN SECURITIES

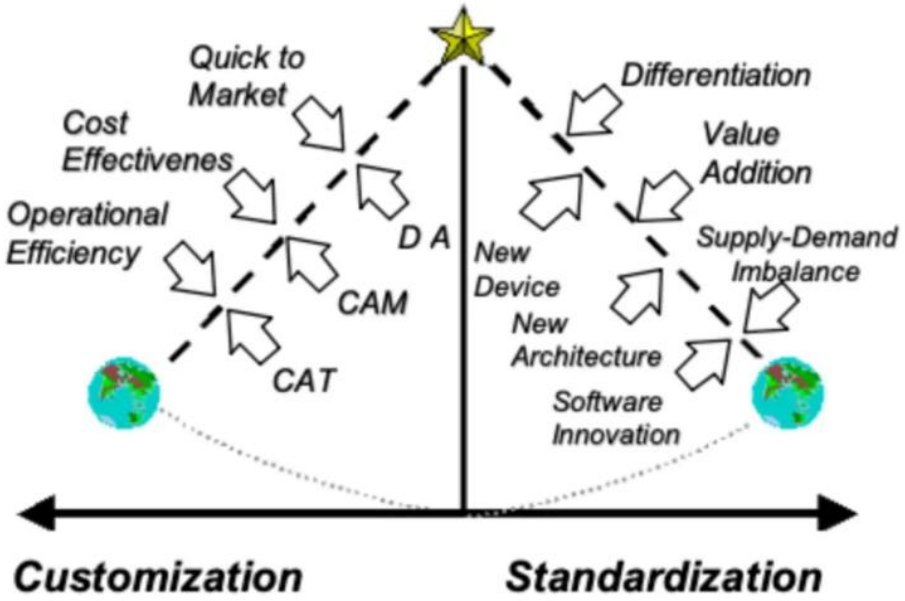
● **牧本摆动每十年波动一次，有望从标准化摆向定制化。**1987年，原日立公司总工程师牧本次生提出牧本摆动，揭露半导体产品发展历程总是在“标准化”与“定制化”之间交替摆动，大概每十年波动一次。牧本摆动背后是性能、功耗和开发效率之间的平衡，当算法发展达到平台期，无法通过进一步创新来推动发展时，就需要依赖于扩大规模来维持进步，这时转向ASIC的开发就变得至关重要。然而十年后，当规模扩张遭遇限制，又会重新聚焦于算法的创新，同时伴随半导体制造技术的进步，一些可编程解决方案在性价比上将会重新获得竞争优势。当前为了满足CSP客户更高性能和更好功能的需求，定制化芯片ASIC的需求持续提升，牧本钟摆从标准化逐渐摆向定制化。

图：标准化制造和定制化应用互相更替



资料来源：土人观芯公众号，国信证券经济研究所整理

图：牧本定律在标准化与定制化之间交替摆动



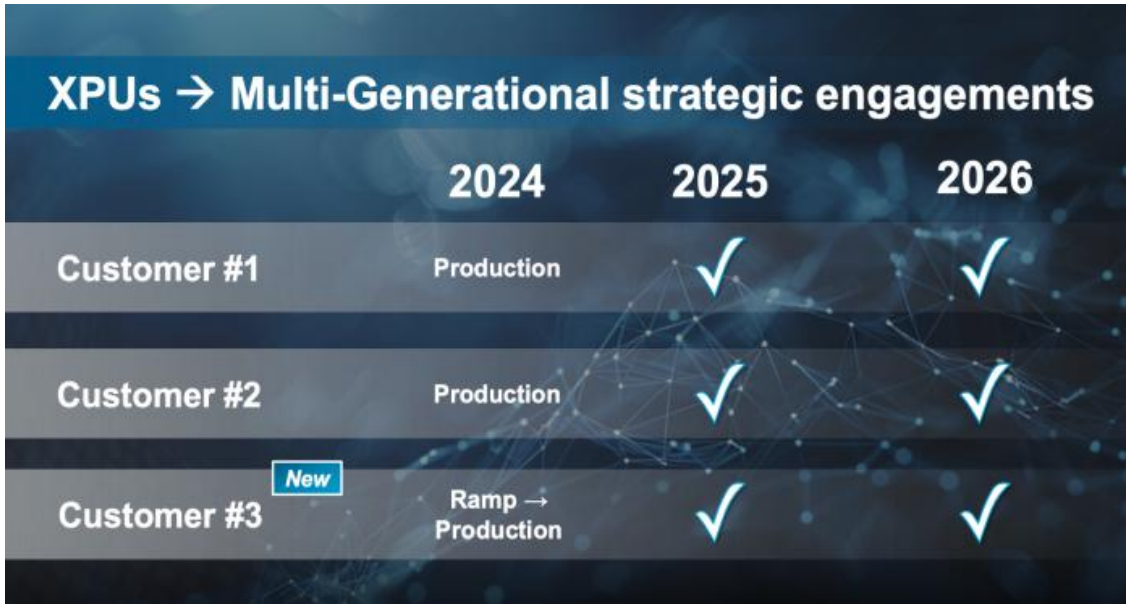
资料来源：土人观芯公众号，国信证券经济研究所整理

牧本定律摆向定制化，关注国产ASIC服务商

国信证券 GUOSEN SECURITIES

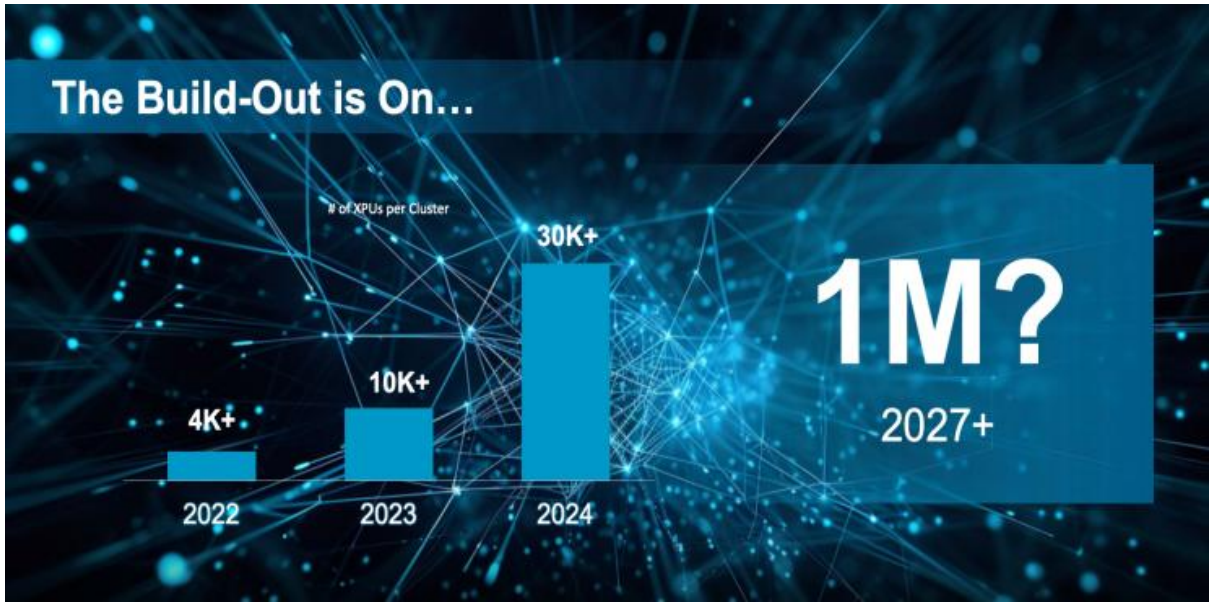
- 在博通2024财年报告中，公司AI业务营收达到约122已美元，同比增长220%。同时，公司预计2027年AI业务可达市场规模为600-900亿美元，客户有望在AI芯片集群中部署100万个芯片，当前公司已开始为三家头部CSP客户提供ASIC。
- 国内具备较强芯片定制服务能力的公司，有望在当前定制化ASIC芯片的趋势中收益。例如，翱捷科技基于丰富的设计经验及雄厚的技术积累，曾为全球领先的人工智能平台公司S、登临科技、美国Moffett等数家知名人工智能技术企业提供先进工艺下的人工智能云端推理超大规模芯片定制服务。

图：博通已为两家头部CSP客户提供ASIC



资料来源：Broadcom官网，国信证券经济研究所整理

图：AI芯片集群有望达到100万张量级



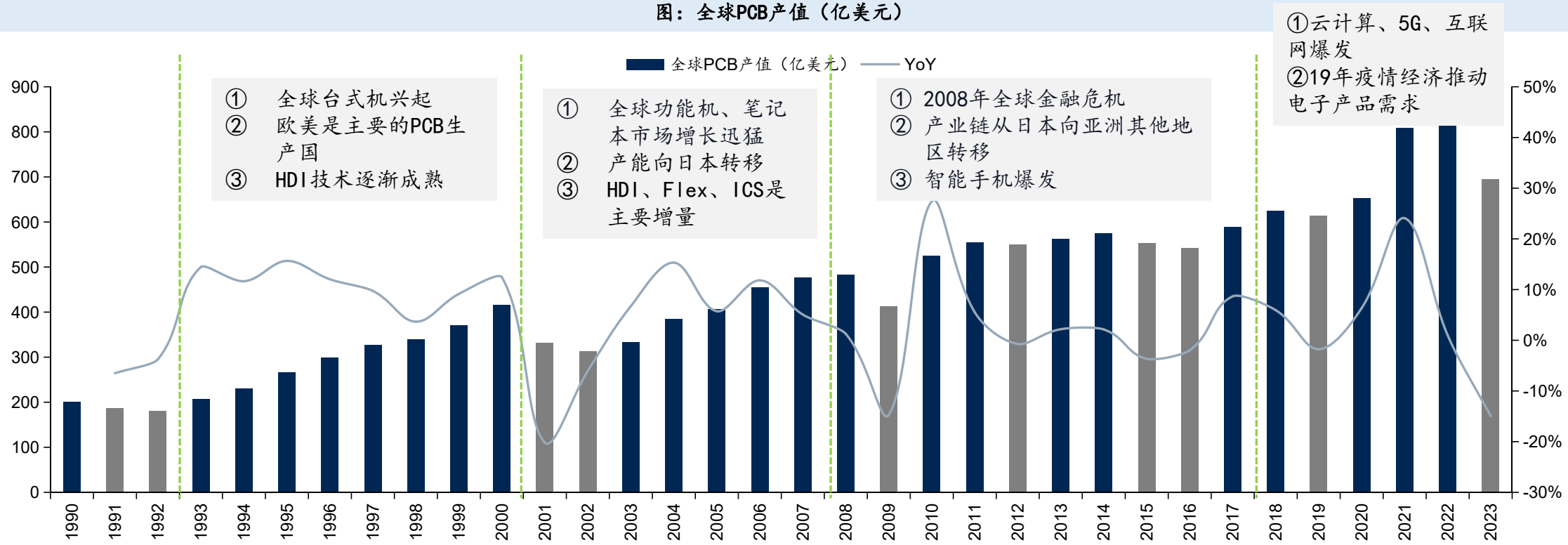
资料来源：Broadcom官网，国信证券经济研究所整理

算力需求是PCB行业的主要增长引擎

国信证券 GUOSEN SECURITIES

- **印制电路板（Printed Circuit Board，PCB）**是指在绝缘基板上，有选择地加工安装孔、连接导线和装配电子元器件的焊盘，以实现电子元器件之间的电气互连的组装板。由于PCB可以实现电路中各元器件之间的电气连接，几乎任何一台电子设备都离不开它，它对电路的电气性能、机械强度和可靠性都起着重要作用，**因此被称为“电子产品之母”**。
- 根据Prismark数据，2023年全球PCB总产值同比下滑14.9%，达到695亿美金规模，Prismark预计2024年全球PCB产值将重回增长，达到730.26亿美金，同比增长5%。

图：全球PCB产值（亿美元）



来源：Prismark，国信证券经济研究所整理

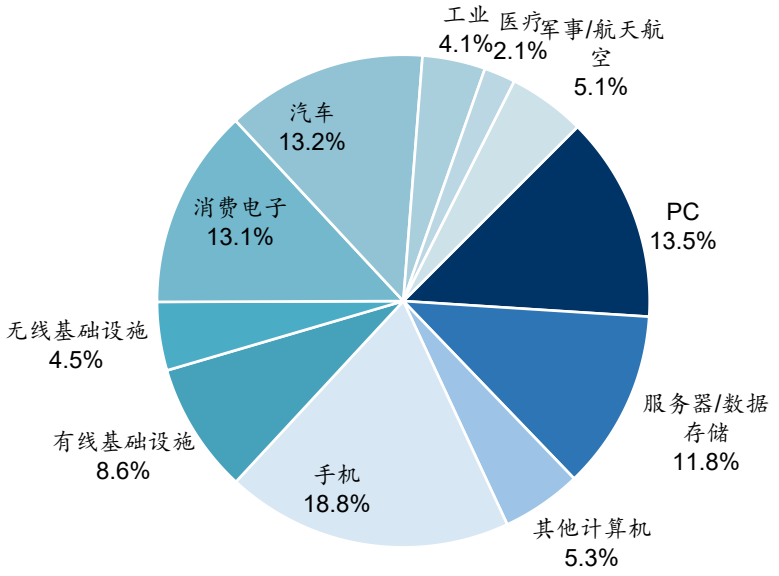
算力需求是PCB行业的主要增长引擎

国信证券 GUOSEN SECURITIES

● **PCB市场下游应用分布广泛，主要涉及计算机、服务器、消费电子、汽车、工业、医疗、军事航天等领域。**根据Prismark 2023年数据，手机占比最大，约为18.8%；其次是个人计算机和消费电子，占比分别约13.5%和13.1%；服务器/数据存储领域的占比也均达12%左右。此外，2023年汽车的占比有所提升，达到13.2%。预计2023-2028年增速最快的是服务器和存储相关PCB，CAGR达到11%，其次为有线通信，CAGR 6%，然后是汽车，CAGR达到5%。

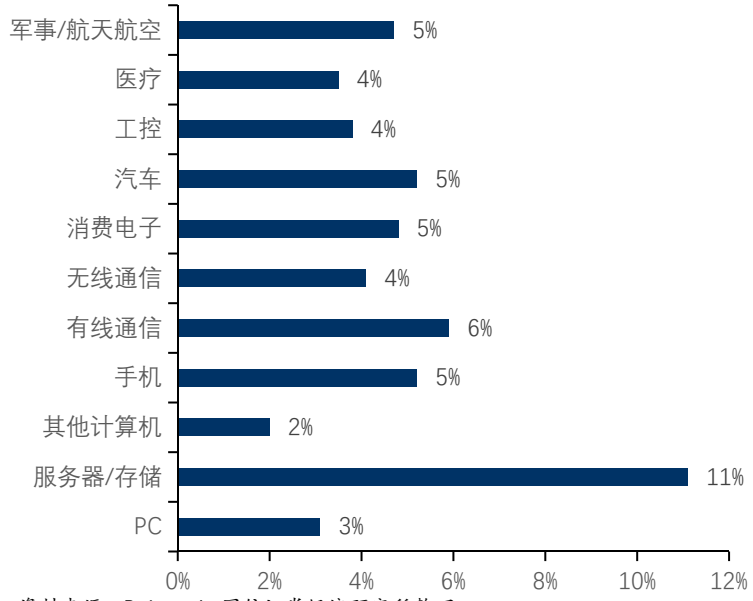
● 从产品种类来看，刚性板的市场规模最大，其中多层板和单双面板的产值占比分别达到36.5%和10.9%；接下来是封装基板，产值占比为21.3%；柔性板和HDI板的产值占比分别为16.9%以及14.4%。

图：2023年全球PCB分下游应用领域产值占比（%）



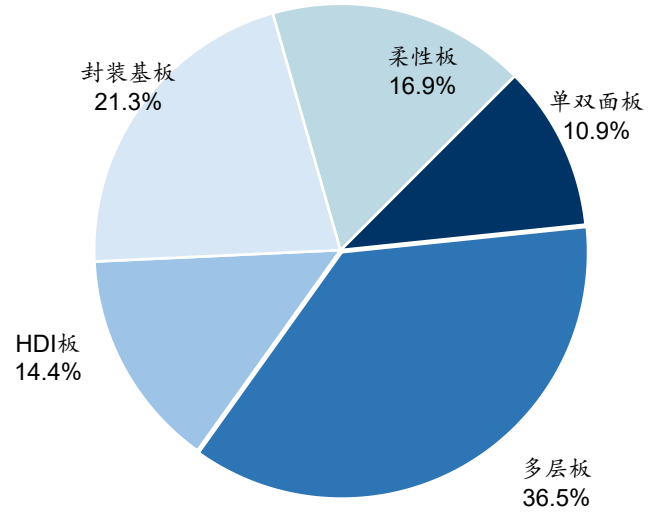
资料来源：Prismark，国信证券经济研究所整理

图：23-28年分应用领域增速预期



资料来源：Prismark，国信证券经济研究所整理

图：2022年全球PCB细分产品的产值占比（%）



资料来源：Prismark，国信证券经济研究所整理

算力需求是PCB行业的主要增长引擎

国信证券 GUOSEN SECURITIES

- **服务器平台的升级会要求PCB板层数增加以及CCL介电损耗降低。**PCB在服务器中的应用主要包括加速板、主板、电源背板、硬盘背板、网卡、Riser卡等，特点主要体现在高层数、高纵横比、高密度及高传输速率。

- **1）PCB板层数增加：**随着服务器平台的演进，服务器PCB持续向更高层板发展，对应于PCIe3.0的Purely服务器平台一般使用8-12层的PCB主板；但Whitley搭载的PCIe4.0总线则要求12-16层的PCB层数；而对于未来将要使用PCIe5.0的Eagle Stream平台而言，PCB层数需要达到16-18层以上。根据Prismark数据，18层以上PCB单价约是12-16层价格的3倍。

图：服务器平台升级要求传输速率提升，Dk、Df下降

英特尔	Purley（Sky Lake）	Purley（Cascade Lake）	Whitley	Eagle Stream
CPU制程	14nm+	14nm++	10nm+	10nm++
PCIe	PCIe3.0	PCIe3.0	PCIe4.0	PCIe5.0
内存	6DDR4	6DDR4	8DDR4	8DDR5
核数	28	28	28	48
传输速率（Gbps）	<28	28	56	112
高速覆铜板类型	Mid-Loss	Mid-loss	Low-Loss	Ultra-Low-Loss
典型Dk值	4.1-4.3	4.1-4.3	3.7-3.9	3.3-3.6
典型Df值	0.008-0.010	0.008-0.010	0.005-0.008	0.002-0.004
对标松下电工产品型号	M4以下	M4一下	M4以上	M6以上

图：服务器升级要求PCB层数增加

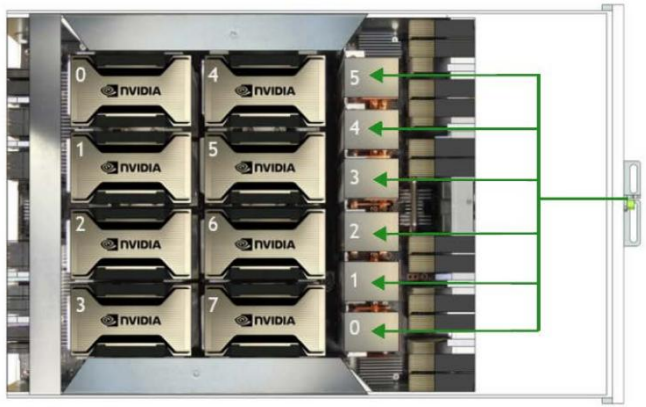
总线标准	对应平台	应用时间	主板层数	CCL材料级别
PCIe3.0	Purley	2017年	10层以下	Mid Loss
PCIe4.0	Whitley	2020年	12-14层	Low Loss
PCIe5.0	Eagle Stream	2022-2023年	16层	Very Low Loss

- **2）高速覆铜板（CCL）介电损耗降低：**服务器主板PCB是由多层导电图形和低介电损耗（Df）的CCL材料压制而成，传输速率要求提高打开Low Loss及以上等级的CCL应用空间。行业内根据CCL的介电损耗Df将CCL划分为STD Loss到Ultra LowLoss六个等级，越高等级损耗越小。PCIe3.0的服务器主板材料以FR4为主，为Mid Loss等级；PCIe4.0主板PCB需升级至Low Loss等级，对应松下M4、生益S7439、联茂IT-958G等材料。

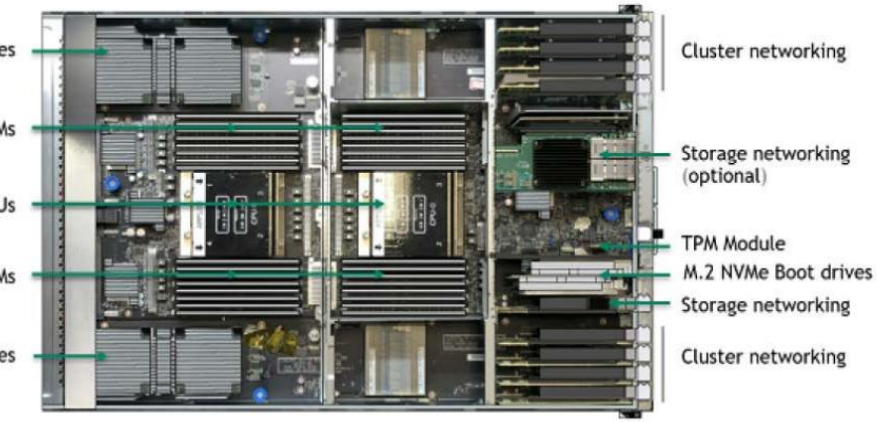
- **新一代英特尔和AMD支持PCIe5.0的服务器平台，主板PCB将继续升级至Ultra Low Loss等级，推动PCB单价进一步提高。**根据Prismark的数据，2019年8-16层PCB板均价约460美元/平方米，18层以上则达到1466美元/平方米，价格增长219%。

来源：Prismark，国信证券经济研究所整理

DGX服务器主要涉及OAM和UBB



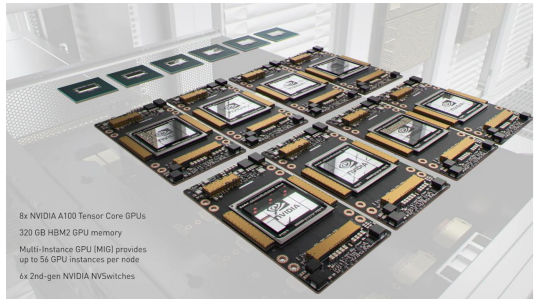
组件	数量	PCB要求
CPU	2	ABF载板（14-16L）
CPU主板	1	多层通孔板（10-14L）
GPU	8	ABF载板（14-16L）
NVSwitch	6	ABF载板

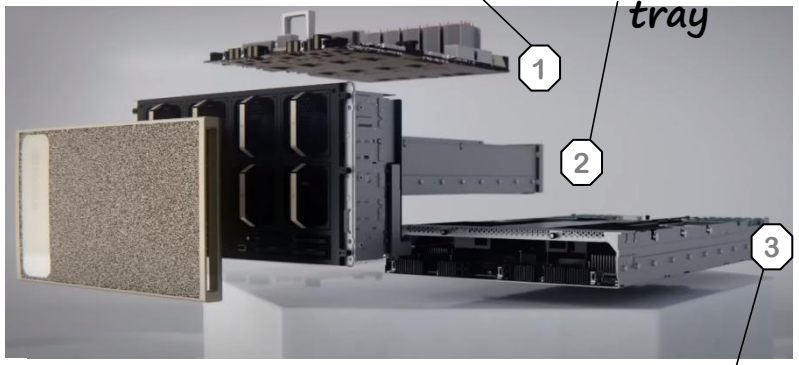


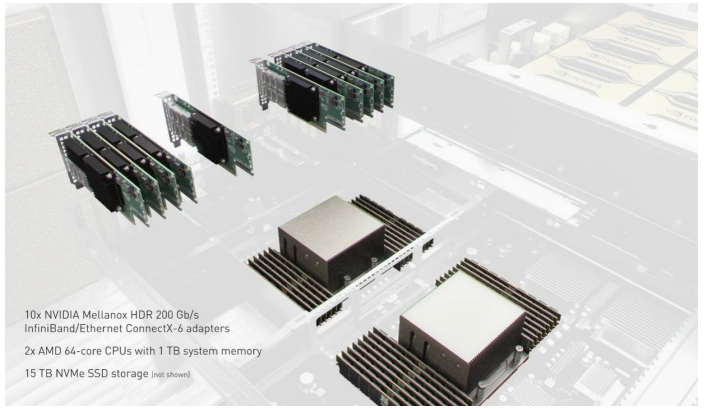
GPU Board tray

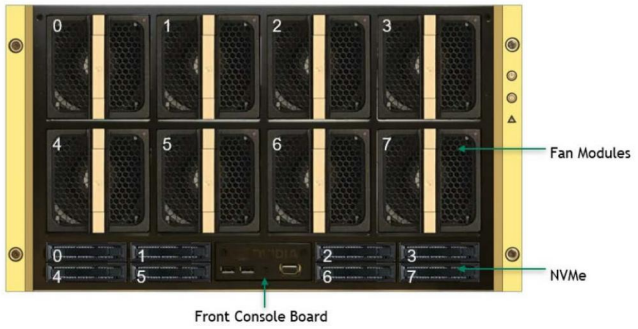
Motherboard tray

Power Supplies

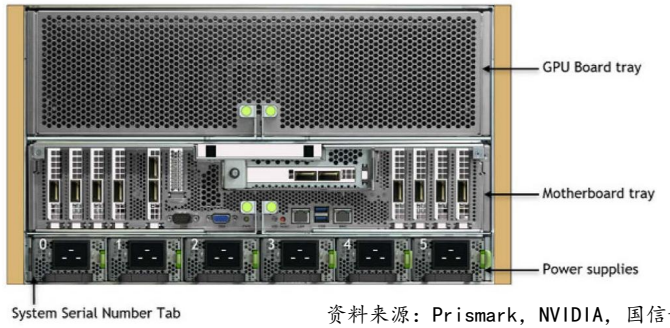








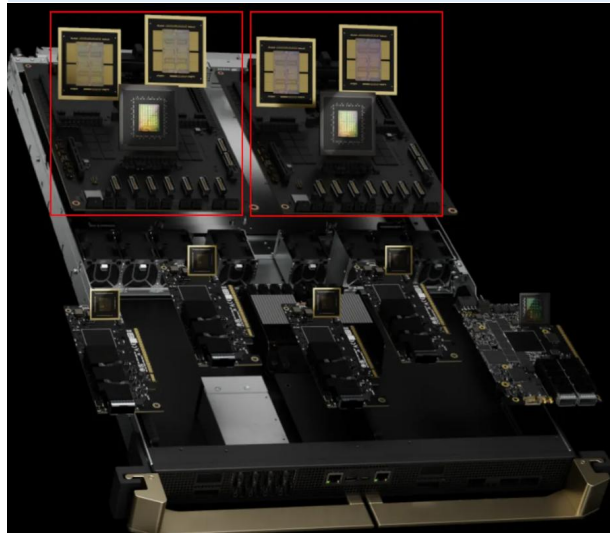
GPU模组主板UBB	1	多层通孔20-26L
GPU加速卡OAM	8	HDI 4N4-6N618L+
内存	32	多层板、BT载板
SSD硬盘	8	多层板



资料来源：Prismark，NVIDIA，国信证券经济研究所整理

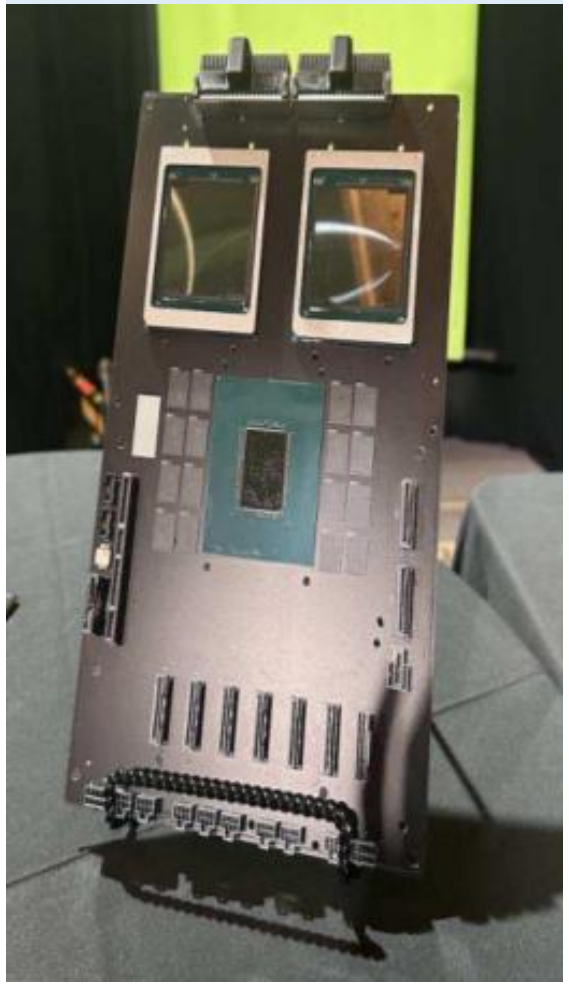
GB200 SuperPOD主要涉及Superchip和Switch board

图：Computer Tray

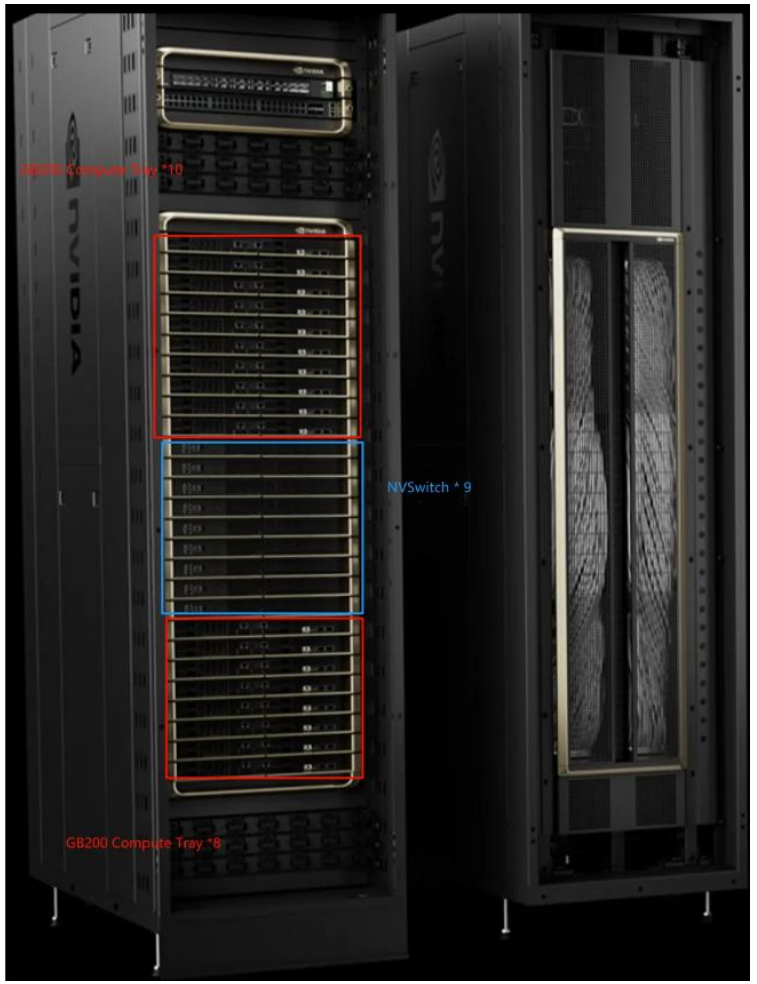


指标	范围	未来5年趋势
层数	低端8L 高端机架 12-16L 最复杂 >22L	范围无变化，但是高层占比会提升
最大层数	24-28	30-34
线间距	4 mil → 3.5 mil	
材料	Mid-Loss到Very-Loss层压板、薄铜	电性能要求更高、需要ultra low-loss, extreme, low-loss层压板、极薄型铜

图：GB200 SuperChip



图：DGX GB200 SuperPOD



资料来源：Prismark，NVIDIA，国信证券经济研究所整理

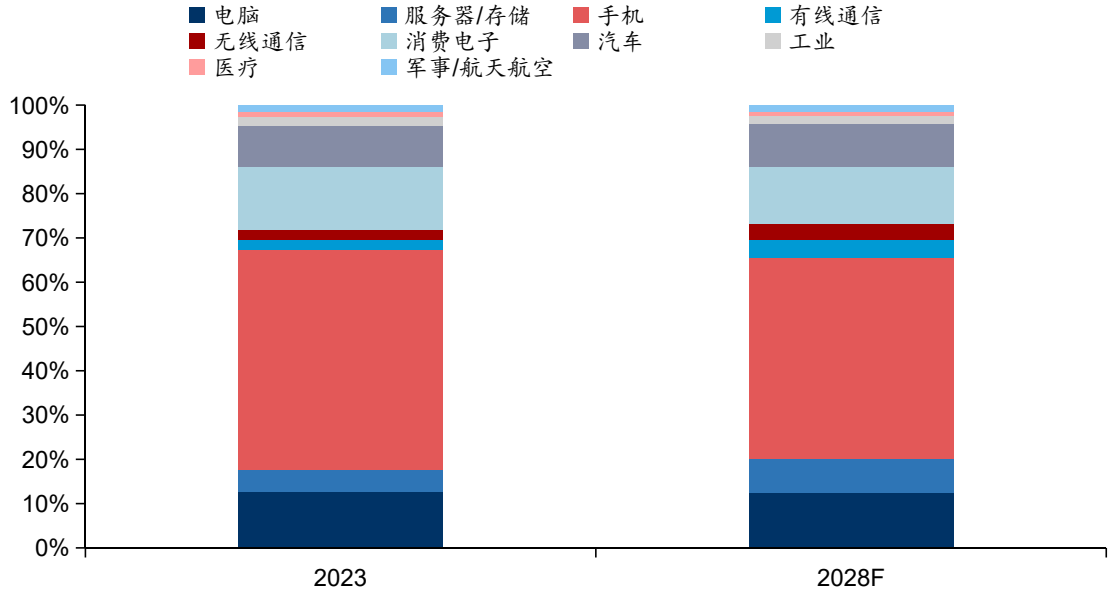
算力需求是PCB行业的主要增长引擎

国信证券 GUOSEN SECURITIES

- **HDI在2020年、2021年增长强劲，分别同比增长9.6%、19.6%，2022年HDI由于中国智能手机需求下滑，市场下滑0.4%。**2023年，由于高存货、下游需求疲软、供大于求和市场竞争加剧导致价格下滑，整体市场下滑严重。1Q24，智能手机的HDI板产量较去年有所改善，由于利润率低，供应能力有所萎缩，低端HDI供应紧张，平均售价从2023年的历史低点回升20%以上。1H24，新的应用领域增速迅猛，卫星通信、汽车智能驾驶和中控板，无线通信、AI GPU模组卡、可穿戴设备、AR/VR等推动了高端HDI的需求。

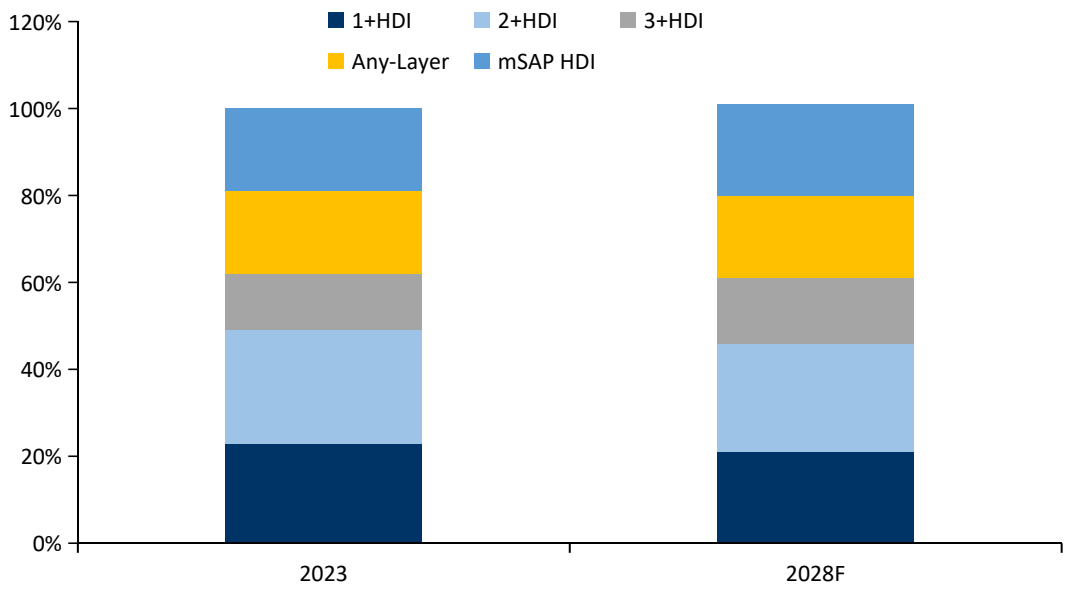
- **预计HDI市场将从2023年的105亿美元增长至2028年的142亿美元，CAGR达到6.2%。**下游具体的应用占比来看，2023年占比最大的智能机份额从50%下滑到45%，增速最快的是有线和无线基建，其次就是服务器和数据存储，CAGR达到16%。**由于高端产品需求增速更快，3+HDI及以上的产品占比预计将从2023年的51%提升到54%。**

图：HDI下游应用



资料来源：Prismark，国信证券经济研究所整理

图：HDI分种类的占比



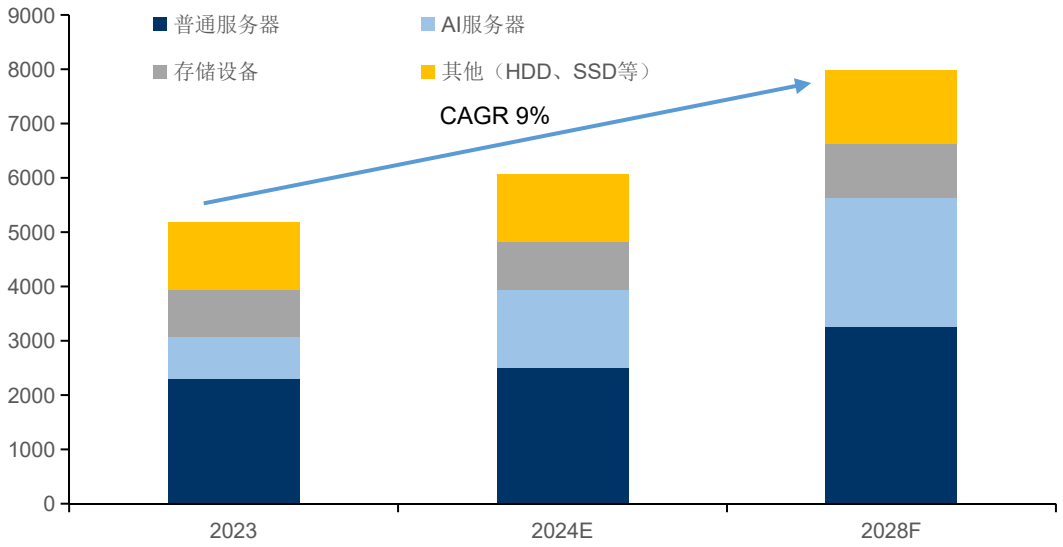
资料来源：Prismark，国信证券经济研究所整理

算力需求是PCB行业的主要增长引擎

国信证券 GUOSEN SECURITIES

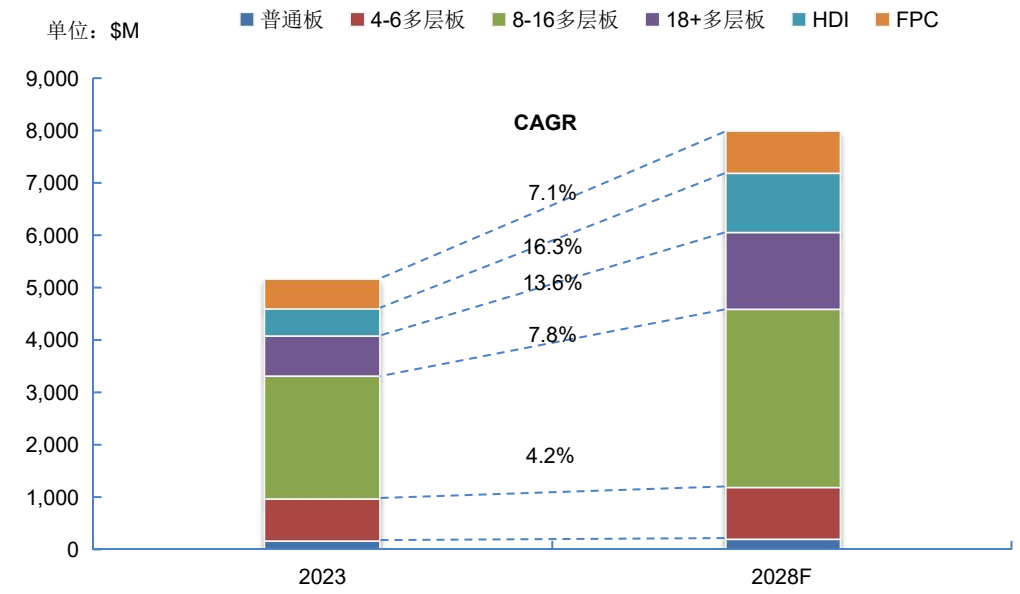
- 根据Prismark数据，2023年全球服务器及相关系统组件的PCB市场规模约为51.77亿美元，预计未来将以9%的增速增长至2028年的79.74亿美元。
- **未来五年AI系统、服务器、存储、网络设备等是PCB需求增长的主要动能。**AI服务器主要涉及3块产品：GPU的基板需要用到20层以上的高多层板，并且使用高速材料；而小型AI加速器模组通常使用HDI来达到高密度互联，通常是4-5阶的HDI；传统的CPU的母板。并且，随着AI服务器升级，GPU主板也将逐步升级为HDI，因此HDI将是未来5年增速最快的PCB，根据Prismark预计，2023-2028年HDI的CAGR将达到16.3%，是增速最快的品类。

图：全球服务器系统及组件PCB市场规模



资料来源：Prismark，国信证券经济研究所整理

图：服务器PCB市场分产品占比



资料来源：Prismark，国信证券经济研究所整理

风险提示

风险提示

国信证券 GUOSEN SECURITIES

1、宏观AI应用推广不及预期。AI技术在应用推广的过程可能面临各种挑战，比如：（1）AI技术需要更多的时间来研发和调试，而且在应用过程中可能会受到数据质量、资源限制和技术能力等因素的制约；（2）AI技术的实施需要更多的资源和资金支持；（3）市场竞争可能也会影响企业在AI应用推广方面的表现。因此，投资者应审慎评估相关企业的技术实力、资金实力以及管理能力，相关企业的AI应用存在推广进度不及预期的风险。

2、AI投资规模低于预期。尽管AI技术在过去几年中受到广泛关注，但AI相关领域的企业投资回报并不总是符合预期。部分企业在AI领域可能缺乏足够的经验和资源，难以把握市场机会。此外，市场竞争也可能会影响企业的投资力度。因此，存在AI领域投资规模低于预期，导致企业相关业务销售收入不及预期的风险。

3、AI服务器渗透率提升低于预期。虽然AI服务器的应用已经较为广泛，但AI服务器渗透率提升的速度存在低于预期的风险，这与企业对AI技术的投资意愿有关，也可能与市场需求和技术进展的速度有关。

4、AI监管政策收紧。由于AI技术的快速发展和广泛应用，监管机构可能会加强对AI技术的监管力度。监管机构可能会制定严格的AI技术使用规定，以保障人们的隐私和数据安全，这些监管政策可能会对企业的业务模式和发展战略造成影响。

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即报告发布日后的6到12个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046　　总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032