信达证券 CINDA SECURITIES

Research and Development Center

AI 产业川流汇聚，云端两旺机遇开启

2025 年 1 月 7 日

证券研究报告

行业研究

行业专题研究（普通）

电子

投资评级 看好

上次评级 看好

莫文宇 电子行业首席分析师

执业编号：S1500522090001

邮 箱：mowenyu@cindasc.com

信达证券股份有限公司

CINDA SECURITIES CO.,LTD

北京市西城区宣武门西大街甲127号金隅大厦B座

邮编：100031

AI 产业川流汇聚，云端两旺机遇开启

2025 年 1 月 7 日

本期内容提要：

- **Blackwell 众多技术突破，整体以机柜形式交货。**GB200 机柜有 NVL36 和 NVL72 两种规格。 GB200 NVL36 配置中，一个机架有 36 个 GPU 和 9 个双 GB200 计算节点（以托盘为单位）。GB200 NVL72 在一个机架中配置了 72 个 GPU / 18 个双 GB200 计算节点，或在两个机架中配置了 72 个 GPU，每个机架上配置了 18 个单 GB200 计算节点。每个 GPU 具有 2080 亿个晶体管，采用专门定制的台积电 4NP 工艺制造。所有 Blackwell 产品均采用双倍光刻极限尺寸的裸片，通过 10 TB/s 的片间互联技术连接成一块统一的 GPU。此外，B 系列还有众多突破，支持 4 位浮点 (FP4) AI。内存可以支持的新一代模型的性能和大小翻倍，同时保持高精度。互联方面，第五代 NVLink 技术实现高速互联。NVIDIA NVLink 交换机芯片能以惊人的 1.8TB/s 互连速度为多服务器集群提供支持。采用 NVLink 的多服务器集群可以在计算量增加的情况下同步扩展 GPU 通信，因此 NVL72 可支持的 GPU 吞吐量是单个 8 卡 GPU 系统的 9 倍。此外，Blackwell 架构在安全 AI、解压缩引擎、可靠性等方面也实现了不同程度的创新和突破。

- **Blackwell 或成推理市场的钥匙，FP4 精度潜力较大。**目前模型参数变大的速度放缓，但模型推理和训练的运算量仍高速增长，尤其在 o1 引入强化学习之后，post scaling law 开始发力。英伟达在发布 H100 架构时，便就 FP8 数据精度做出一定讨论。业界曾长期依赖 FP16 与 FP32 训练，但这种高精度的运算，在大模型 LLM 中受到了一定阻碍：由于模型参数等因素导致运算骤升，可能导致数据溢出。英伟达提出的 FP8 数据精度因为占用更少的比特，能提供更多运算量。以 NVIDIA H100 Tensor Core GPU 为例，相较 FP16 和 BF16，FP8 的峰值性能能够实现接近翻倍。FP4 精度是 FP8 的继承和发展，对推理市场的打开有重要推动。GB200 推出了 FP4，FP4 支持由于降低了数据精度，性价比相比 H100 几乎倍增。根据 Semianalysis 的数据，GB200 NVL72 在 FP4 精度下，FLOPS 相比 H100 可以最高提高 405%（注：H100 最低以 FP8 计算），由此带来性价比提升。目前，FP4 的运算已经可以在大模型运算中广泛应用，且已有研究表明网络可以使用 FP4 精度进行训练而不会有显著的精度损失。此外，由于模型推理中不需要对模型参数进行更新，相对训练对于精度的敏感性有所下降，因此 B 系列相对于训练，在推理领域会更有优势。B 系列引入 FP4 精度后，大模型在云侧和端侧的协同都有望实现跃升，这也是我们看好接下来的端侧市场的原因之一。

- **AI 产业川流汇聚，2025 年有望云端两旺。**我们认为， B 系列的推出有望打开推理市场，各类 AI 终端有望掀起持续的机遇。此外，AI 产业的闭环有望刺激云厂商资本开支，云端共振共同发展。建议关注英伟达产业链传统的核心厂商，如 ODM、PCB 厂商等。此外，B 系列带来的新兴赛道如铜连接、AEC 赛道也值得关注。

- **风险因素：**宏观经济下行风险；下游需求不及预期风险；中美贸易摩擦加剧风险。

目 录

表 目 录

图 目 录

Blackwell 众多技术突破，整体以机柜形式交货

GB200 机柜有 NVL36 和 NVL72 两种规格。 GB200 NVL36 配置中，一个机架有 36 个 GPU 和 9 个双 GB200 计算节点（以托盘为单位）。GB200 NVL72 在一个机架中配置了 72 个 GPU / 18 个双 GB200 计算节点，或在两个机架中配置了 72 个 GPU，每个机架上配置了 18 个单 GB200 计算节点。

图 1：GB200 NVL72 机柜正面

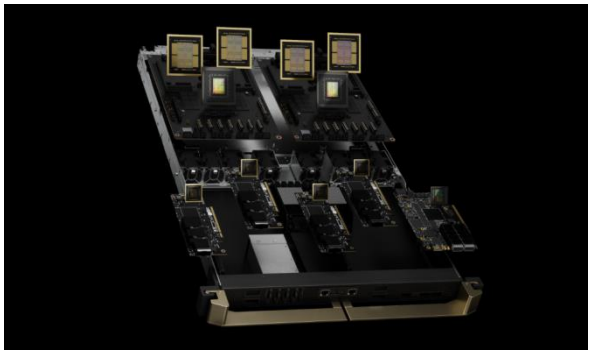
资料来源：英伟达官网，信达证券研发中心

图 2：GB200 NVL72 机柜背面

资料来源：英伟达官网，信达证券研发中心

- **计算托盘：**每一个计算托盘有两个 NVIDIA GB200 Grace Blackwell 超级芯片。每个超级芯片将两个高性能 NVIDIA Blackwell Tensor Core GPU 和 NVIDIA Grace CPU 与 NVLink 芯片到芯片 （C2C） 接口连接起来，可提供 900 GB/s 的双向带宽。借助 NVLink-C2C，应用程序可以一致地访问统一的内存空间。这简化了编程，并支持万亿参数 LLM、用于多模态任务的 transformer 模型、用于大规模仿真的模型以及用于 3D 数据的生成模型的更大内存需求。
- **交换托盘：**NVIDIA GB200 NVL72 引入了第五代 NVLink，它可以在单个 NVLink 域中连接多达 576 个 GPU，总带宽超过 1 PB/s，快速内存为 240 TB。每个 NVLink 交换机托盘提供 144 个 100 GB 的 NVLink 端口，因此这 9 台交换机完全连接了 72 个 Blackwell GPU 上每个 GPU 上的 18 个 NVLink 端口中的每一个。每个 GPU 的革命性 1.8 TB/s 双向吞吐量是 PCIe Gen5 带宽的 14 倍以上，为当今最复杂的大型模型提供无缝高速通信。

图 3：GB200 机柜



资料来源：英伟达官网，信达证券研发中心

图 4：GB200 机柜背面

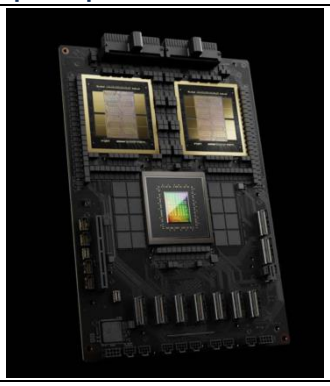
资料来源：英伟达官网，信达证券研发中心

Blackwell 架构实现了较多的技术突破:

- **GPU 工艺难度和晶体管数量上升。**每个 GPU 具有 2080 亿个晶体管，采用专门定制的台积电 4NP 工艺制造。所有 Blackwell 产品均采用双倍光刻极限尺寸的裸片，通过 10 TB/s 的片间互联技术连接成一块统一的 GPU。
- **第二代 Transformer 引擎及针对推理推出 FP4 数据精度。**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 (LLM) 和专家混合模型 (MoE) 的推理和训练。为了强效助力 MoE 模型的推理 Blackwell Tensor Core 增加了新的精度 (包括新的社区定义的微缩放格式)，可提供较高的准确性并轻松替换更大的精度。Blackwell Transformer 引擎利用称为微张量缩放的细粒度缩放技术，优化性能和准确性，支持 4 位浮点 (FP4) AI。这将内存可以支持的新一代模型的性能和大小翻倍，同时保持高精度。
- **第五代 NVLink 技术实现高速互联。**第五代 NVIDIA® NVLink® 可扩展至 576 个 GPU，为万亿和数万亿参数 AI 模型释放加速性能。NVIDIA NVLink 交换机芯片可在一个有 72 个 GPU 的 NVLink 域 (NVL72) 中实现 130TB/s 的 GPU 带宽，并通过 NVIDIA SHARP™ 技术对 FP8 的支持实现 4 倍于原来的带宽效率。NVIDIA NVLink 交换机芯片能以惊人的 1.8TB/s 互连速度为多服务器集群提供支持。采用 NVLink 的多服务器集群可以在计算量增加的情况下同步扩展 GPU 通信，因此 NVL72 可支持的 GPU 吞吐量是单个 8 卡 GPU 系统的 9 倍。

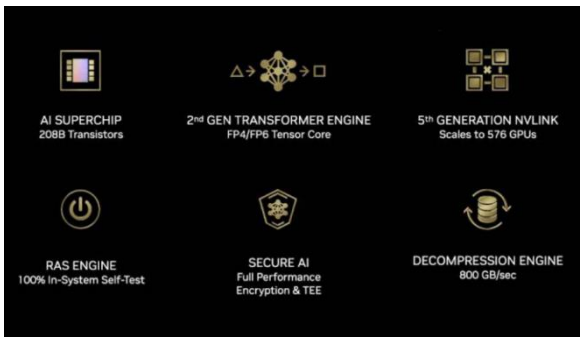
此外，Blackwell 架构在安全 AI、解压缩引擎、可靠性等方面也实现了不同程度的创新和突破。

图 5: GB200 Superchip



资料来源：英伟达官网，信达证券研发中心

图 6: Blackwell 的技术突破

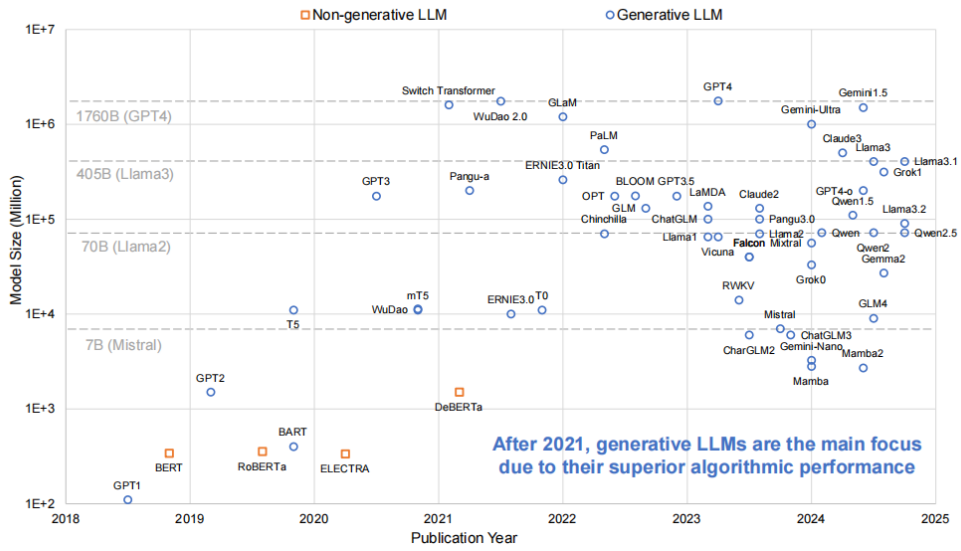


资料来源：英伟达官网，信达证券研发中心

Blackwell 或成推理市场的钥匙，FP4 精度潜力较大

目前模型的参数变大的速度放缓，但模型推理和训练的运算量仍高速增长。由于高质量训练语料的限制，目前模型参数变大的速度正在放缓。但是，模型训练和推理的运算量却在上升，尤其在 o1 引入强化学习之后，post-scaling law 开始发力。

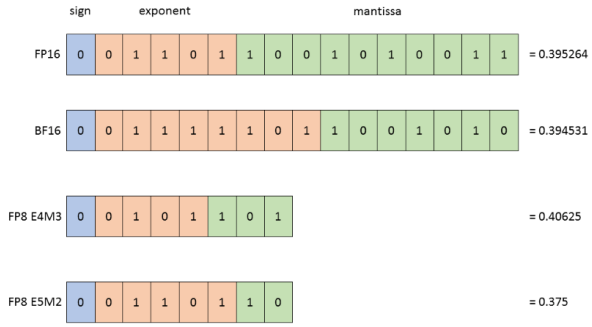
图 7：全球服务器出货按价格带分布（万台）



资料来源：Jinhao Li et al.《Large Language Model Inference Acceleration: A Comprehensive Hardware Perspective》，信达证券研发中心

英伟达在发布 Hopper 架构时，便就 FP8 数据精度做出一定讨论。业界曾长期依赖 FP16 与 FP32 训练，但这种高精度的运算，在大模型 LLM 中受到了一定阻碍：由于模型参数等因素导致运算骤升，可能导致数据溢出。英伟达提出的 FP8 数据精度因为占用更少的比特，能提供更多运算量。以 NVIDIA H100 Tensor Core GPU 上为例，相较 FP16 和 BF16，FP8 的峰值性能能够实现接近翻倍。

图 8：四种数据精度



资料来源：英伟达官网，信达证券研发中心

图 9：英伟达 H100 相对 A100 有较大峰值性能提升（TFLOPS）

NVIDIA H100 specifications (vs. NVIDIA A100)

Data type	H100-SXM5 (TFLOPS)	A100-SXM4 (TFLOPS)	Difference
TF32	494	156	3.2x
BF16	989	312	3.2x
FP16	989	312	3.2x
FP8	1979	-	6.3x (vs BF16)
Bandwidth (GB/s)	3350	2039	1.6x

Table 1: FLOPS and memory bandwidth comparison between the NVIDIA H100 and NVIDIA A100. While there are 3x-6x more total FLOPS, real-world models may not realize these gains.

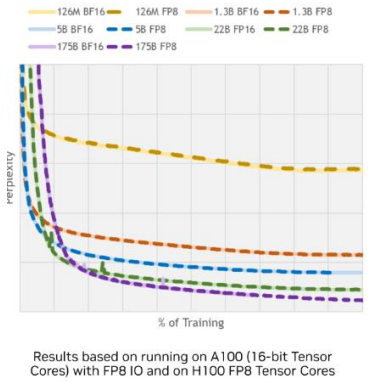
资料来源：英伟达官网，信达证券研发中心

训练方面：在不同规模的 GPT 模型上使用 BF16 与 FP8 进行训练的 loss（损失值）曲线上，以困惑度 PPL（Perplexity）为度量指标。观察 PPL 曲线走势，可见随着训练进程，FP8 与 BF16 的曲线几乎完全吻合，表明两者收敛性并无显著差异。

推理方面：单纯启用 FP8 会由于 batch size 提升有限，以及 KV cache 的影响，导致性能提升并不显著。然而，一旦将 KV cache 也转换至 FP8，通过减半其内存消耗，模型吞

吐量可以相较 FP16 提升约两倍左右，这是一个理想的性能提升幅度。此外，从英伟达官方展示的资料来看，在运算量越大的任务中，FP8 的性能提升幅度越接近上限。

图 10：训练：在不同规模的 GPT 模型上使用 BF16 与 FP8 进行训练的 loss



资料来源：英伟达官网，信达证券研发中心（注：同色曲线代表相同模型规模，实线代表 BF16，虚线为 FP8）

图 11：推理：使用 Tensor-LLM 实现 FP8 推理的性能

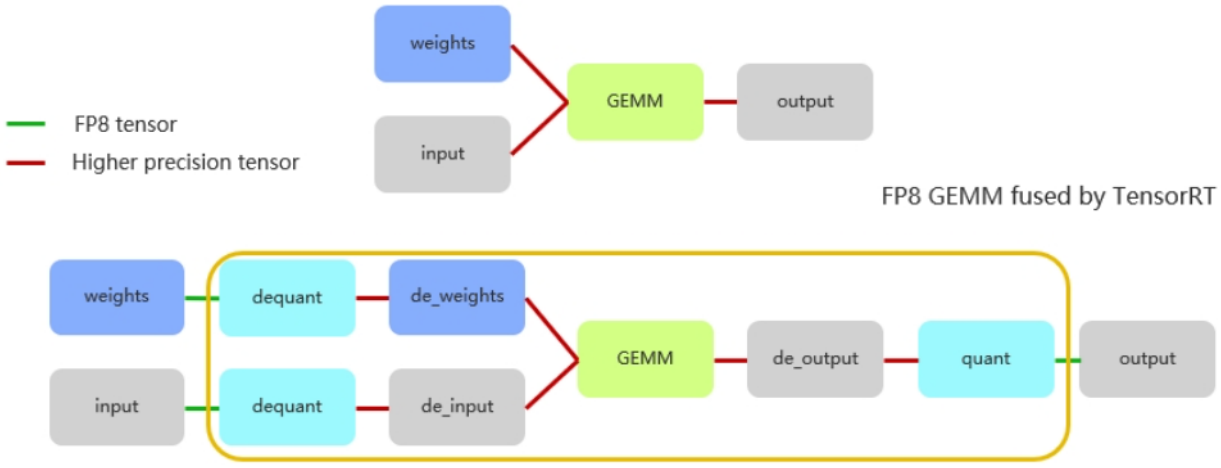
Batch size	FP16 (max bs 75)		FP8 (max bs 85)		FP8 with FP8 KV cache (max bs 169)	
	Memory (GB)	Tokens per sec	Memory (GB)	Tokens per sec	Memory (GB)	Tokens per sec
1	38.61	111.44	33.42	131.74	33.15	130.16
8	42.46	682.68	37.27	872.64	35.09	927.77
64	73.25	1772.07	68.06	2154.42 (1.21x)	50.51	2878.15 (1.62x)
MAX	79.22	1817.32	79.52	2309.20 (1.27x)	79.41	3651.05 (2.00x)

GPT-J 6b, input length 1024, output length 256, CUDA 12.2, Driver 535.104.05, TensorRT 9.1 on *H100

资料来源：英伟达官网，信达证券研发中心（注：max 值指的是在设定输入为 1,024，输出为 256，模型为 GPT-J 6B，所能使用的最大 batch size）

英伟达用 FP8 进行低精度训练取得较好的性能，模型量化是其中的重要手段。模型量化是一种深度学习的优化技术，目前已经取得了较为可观的进展。模型量化可以将神经网络中原本高精度数据运算转换至低精度，这样的优点主要有：（1）减少内存带宽和存储空间；（2）提高系统吞吐量；（3）降低系统延时等等。

图 12：FP8 推理过程



资料来源：英伟达官网，信达证券研发中心

FP4 精度是 FP8 的继承和发展，对推理市场的打开有重要推动。GB200 推出了 FP4，FP4 支持由于降低了数据精度，性价比相比 H100 几乎倍增。根据 Semianalysis 的数据，GB200 NVL72 在 FP4 精度下，FLOPS 相比 H100 可以提高 405%（注：H100 最低以 FP8 计算），由此带来性价比提升。

单位美元运算量：目前 B 的价格仍为公布，但是即便假设 GB200 NVL72 中的单张 GPU 价格为 3.5 万美元，单位美元可提供的算力相对 H100 仍是倍增。

存储：FP8 精度下的运算对于内存的耗用相对较小，因此可以节省存储。或者说单位存储可以提供的运算能力倍增。

功耗：GB200 单张 GPU 的功耗相对 H100 提升 71%，算力提升 405%，如此测算单位算力所耗费的能量大幅减少。

图 13：英伟达产品算力对比

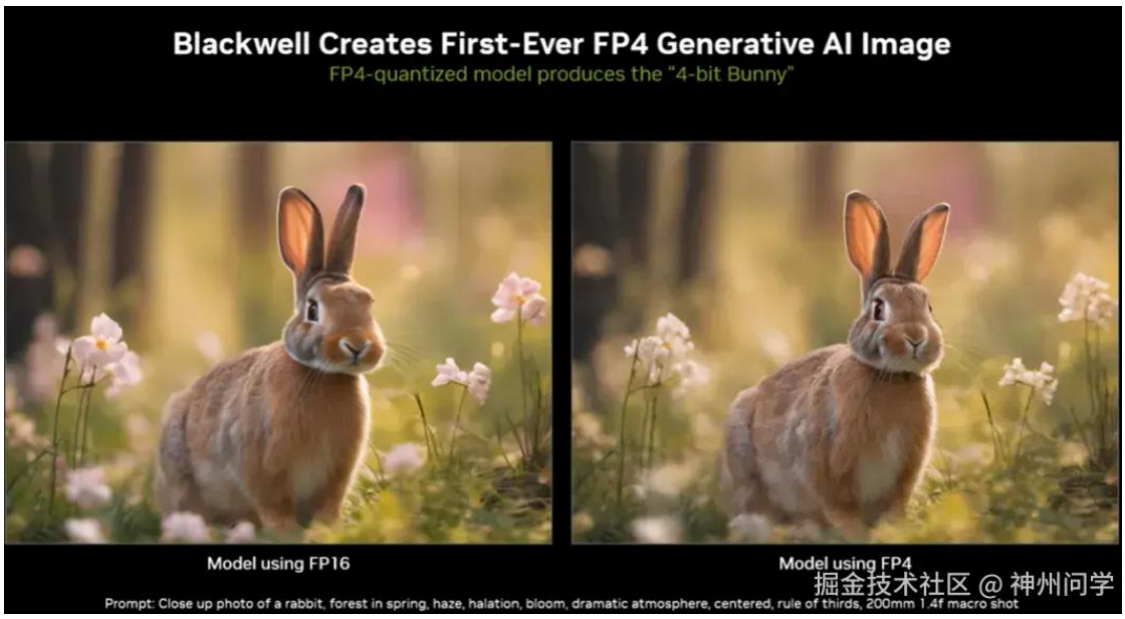
Blackwell vs Hopper Basic Specifications					
	H100	H200	B100	B200	GB200 NVL72
Price[1]	$24,000	$24,000	$30,000		
Watts Per GPU	700	700	700	1,000	1,200
All-in System Watts Per GPU	1,275	1,275	1,275	1,788	1,667
NVLink Bandwidth (Unidirectional - GB/s)	450	450	900	900	900
Memory Capacity (GB)	80GB	141GB	Up to 192GB[3]	Up to 192GB	192GB
Memory Bandwidth (GB/s)	3,352	4,800	Up to 8000[3]	Up To 8,000	8,000
Memory Bandwidth Improvement	*0%*	*43%*	*79%*	*139%*	*139%*
TF32 TFLOPS[2]	495	495	900	1,100	1,250
TF32 Improvement	*0%*	*0%*	*77%*	*127%*	*153%*
FP16/BF16 TFLOPS[2]	989	989	1,750	2,250	2,500
FP16/BF16 Improvement	*0%*	*0%*	*77%*	*127%*	*153%*
FP8 / FP6 / Int8 TFLOPS[2]	1,979	1,979	3,500	4,500	5,000
FP8 / FP6 / Int8 Improvement	*0%*	*0%*	*77%*	*127%*	*153%*
FP4 TFLOPS[2]	1,979	1,979	7,000	9,000	10,000
FP4 Improvement	*0%*	*0%*	*254%*	*355%*	*405%*

1. Full ASP and volume details available in Accelerator Model
2. All FLOPS are dense
3. B100 memory specs to be finalized - we assume 168GB and 6,000 GB/s

资料来源：Nvdia，Semianalysis，信达证券研发中心

目前，FP4 的运算已经可以大模型运算中广泛应用，且已有研究表明网络可以使用 FP4 精度进行训练而不会有显著的精度损失。此外，由于模型推理中不需要对模型参数进行更新，相对训练对于精度的敏感性有所下降，因此 B 系列相对于训练，在推理领域会更有优势。**此外，B 系列引入 FP4 精度后，大模型在云侧和端侧的协同都有望实现跃升，这也是我们看好接下来的端侧市场的原因之一。**

图 14：FP16 和 FP4 精度下生成的图片对比



资料来源：稀土掘金，神州问学，信达证券研发中心

AI 产业川流汇聚，2025 年有望云端两旺。我们认为， B 系列的推出有望打开推理市场，各类 AI 终端有望掀起持续的机遇。此外，AI 产业的闭环有望刺激云厂商资本开支，云端共振共同发展。建议关注英伟达产业链传统的核心厂商，如 ODM、PCB 厂商等。此外，B 系列带来的新兴赛道如铜连接、AEC 赛道也值得关注。

表 1：建议关注

			净利润（亿元）			PE		
股票代码	股票简称	总市值（亿元）	2024E	2025E	2026E	2024E	2025E	2026E
601138	工业富联	4,106.72	237.41	300.72	341.24	17.30	13.66	12.03
002463	沪电股份	765.14	25.44	36.37	46.25	30.08	21.04	16.54
300476	胜宏科技	350.60	11.78	17.48	21.63	29.75	20.06	16.21
002916	深南电路	603.55	20.77	25.36	30.41	29.07	23.80	19.84
002130	沃尔核材	300.86	9.30	12.12	14.66	32.35	24.82	20.52
600183	生益科技	550.26	18.80	23.72	28.60	29.26	23.20	19.24
688183	生益电子	302.28	3.01	6.86	9.36	100.48	44.05	32.31

资料来源：ifind，信达证券研发中心（除工业富联、沪电股份外为 ifind 一致预期，截至 2025 年 1 月 6 日）

风险因素

宏观经济下行风险；

下游需求不及预期风险；

中美贸易摩擦加剧风险。

研究团队简介

莫文宇，电子行业分析师，S1500522090001。毕业于美国佛罗里达大学，电子工程硕士，2012-2022 年就职于长江证券研究所，2022 年入职信达证券研发中心，任副所长、电子行业首席分析师。

郭一江，电子行业研究员。本科兰州大学，研究生就读于北京大学化学专业。2020 年 8 月入职华创证券电子组，后于 2022 年 11 月加入信达证券电子组，研究方向为光学、消费电子、汽车电子等。

王义夫，电子行业研究员。西南财经大学金融学士，复旦大学金融硕士，2023 年加入信达证券电子组，研究方向为存储芯片、模拟芯片等。

李星全，电子行业研究员。哈尔滨工业大学学士，北京大学硕士。2023 年加入信达证券电子组，研究方向为服务器、PCB、消费电子等。

分析师声明

负责本报告全部或部分内容的每一位分析师在此申明，本人具有证券投资咨询执业资格，并在中国证券业协会注册登记为证券分析师,以勤勉的职业态度,独立、客观地出具本报告；本报告所表述的所有观点准确反映了分析师本人的研究观点；本人薪酬的任何组成部分不曾与，不与，也将不会与本报告中的具体分析意见或观点直接或间接相关。

免责声明

信达证券股份有限公司(以下简称“信达证券”)具有中国证监会批复的证券投资咨询业务资格。本报告由信达证券制作并发布。

本报告是针对与信达证券签署服务协议的签约客户的专属研究产品，为该类客户进行投资决策时提供辅助和参考，双方对权利与义务均有严格约定。本报告仅提供给上述特定客户，并不面向公众发布。信达证券不会因接收人收到本报告而视其为本公司的当然客户。客户应当认识到有关本报告的电话、短信、邮件提示仅为研究观点的简要沟通，对本报告的参考使用须以本报告的完整版本为准。

本报告是基于信达证券认为可靠的已公开信息编制，但信达证券不保证所载信息的准确性和完整性。本报告所载的意见、评估及预测仅为本报告最初出具日的观点和判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会出现不同程度的波动，涉及证券或投资标的的历史表现不应作为日后表现的保证。在不同时期，或因使用不同假设和标准，采用不同观点和分析方法，致使信达证券发出与本报告所载意见、评估及预测不一致的研究报告，对此信达证券可不发出特别通知。

在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，也没有考虑到客户特殊的投资目标、财务状况或需求。客户应考虑本报告中的任何意见或建议是否符合其特定状况，若有必要应寻求专家意见。本报告所载的资料、工具、意见及推测仅供参考，并非作为或被视为出售或购买证券或其他投资标的的邀请或向人做出邀请。

在法律允许的情况下，信达证券或其关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能会为这些公司正在提供或争取提供投资银行业务服务。

评级说明

投资建议的比较标准	股票投资评级	行业投资评级
本报告采用的基准指数 ：沪深 300 指数（以下简称基准）； 时间段：报告发布之日起 6 个月内。	**买入：**股价相对强于基准 15%以上；	**看好：**行业指数超越基准；
	增持：股价相对强于基准 5%～15%；	**中性：**行业指数与基准基本持平；
	持有：股价相对基准波动在±5% 之间；	**看淡：**行业指数弱于基准。
	卖出：股价相对弱于基准 5%以下。	

风险提示

证券市场是一个风险无时不在的市场。投资者在进行证券交易时存在赢利的可能，也存在亏损的风险。建议投资者应当充分深入地了解证券市场蕴含的各项风险并谨慎行事。

本报告中所述证券不一定能在所有的国家和地区向所有类型的投资者销售，投资者应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专业顾问的意见。在任何情况下，信达证券不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任，投资者需自行承担风险。