国投证券 SDIC SECURITIES

2024 年 01 月 17 日
电子

行业专题

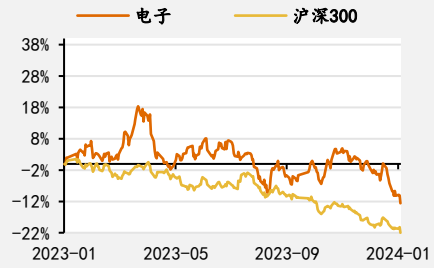
证券研究报告

AI 浪潮势不可挡，昇腾发力铸造国产算力基石

投资评级 领先大市-A
维持评级

首选股票	目标价（元）	评级

行业表现



资料来源：Wind 资讯

升幅%	1M	3M	12M
相对收益	-7.9	-0.3	9.4
绝对收益	-11.2	-11.6	-12.6

马良 分析师
SAC 执业证书编号：S1450518060001
maliang2@essence.com.cn

朱思 分析师
SAC 执业证书编号：S1450523090002
zhusi1@essence.com.cn

相关报告

CES 展会新品纷呈，AI 与显示交互成为焦点	2024-01-14
高端国产替代系列--光刻胶：半导体制造核心材料，国产替代突围在即	2024-01-10
高通推出第二代骁龙 XR2+平台，原生鸿蒙生态有望迎来快速发展	2024-01-07
华为、小米新车相继发布，全球首款商务 AI PC 亮相	2024-01-01
苹果 MR 发售在即，美光业绩超预期	2023-12-24

从“性能指标”到“性能密度指标”，海外高端芯片进口受限范围扩大：

2023 年 10 月 17 日，美国商务部出台了出口管制清单的 ECNN 3A090 和 4A090 要求，以进一步限制高性能 AI 芯片的出口，同时将 13 家中国公司列入实体清单。在新规发布之前，超过旧规性能指标限制的芯片仅为英伟达 A100，但当加入性能密度指标后，新规不仅限制了厂商出略低于性能标准的芯片以规避限制情况，同时针对数据中心芯片与非数据中心芯片进行了不同的限制约定，使更多的英伟达芯片受到禁令限制。修改后的出国管制设计产品包括但不限于：英伟达 A100、A800、H100、H800、L40、L40S 以及 RTX 4090 产品。实际上，任何集成了一个或多个及以上的芯片的系统，包括但不限于英伟达 DGX、HGX 系统，都在新规涵盖范围之内。

全球科技巨头纷纷布局算力芯片，AI 浪潮势不可挡：

AMD 推出的“MI 系列+Infinity Fabric+ROCm 平台”，性能强劲，成为英伟达全球范围内最强劲的对手。特斯拉自研 Dojo 超算服务器助力自动驾驶，在芯片间互连技术上独具特色，具备强大的扩展性。Intel 推出 GPU 系列芯片和 oneAPI 开发平台，不断完善其 AI 数据中心布局。Google 推出 Cloud TPU 解决方案，在机器学习领域。Meta 布局自研 AI 生态，2020 年正式推出第一代 MTIA 方案，侧重于处理低/中复杂度模型。英伟达在 GPU 领域深耕数十年，根据 Jon Peddie Research（JPR）报告显示，2023 年 Q1，英伟达 GPU 市场份额达到 84%，Q2 达到 80%，但全球各大科技巨头结合其自身优势纷纷布局算力芯片。

完整的芯片生态“软件+硬件”缺一不可，华为推出全系列昇腾解决方案：

一个完整的芯片生态中仅有硬件芯片是不够的，软件系统开发的便捷性以及现有系统的迁移难度同样至关重要。现阶段英伟达通过其“芯片+NVLINK+CUDA”构建产品高护城河，在 GPU 市场长期占据全球市场领导地位。华为推出全系列昇腾解决方案，包括自研的昇腾芯片+高速芯片间接口+AI 计算架构，从软硬件上和主流厂商对标，昇腾芯片基于自研的达芬奇架构，功耗低，产品性能优异，从官方披露的数据来看，2019 年推出的昇腾 910，在半精度 FP16 运算速度可达

320TFLOPS，整型 INT8 算力可达 640TOPS，英伟达 A100 的 FP16 运算性能为 312TFLOPS。

相关公司：兴森科技、新益昌、天承科技、德邦科技、华海诚科、英维克、飞荣达、思泉新材、恒铭达、华丰科技、沪电股份、世运电路、胜宏科技、方正科技。

风险提示：宏观因素波动影响下游需求；市场开拓不及预期；行业竞争加剧；

内容目录

图表目录

1. 美国加强限制规则，海外高性能芯片进口受限

1.1. 从“性能指标”到“性能密度指标”，英伟达高端芯片进口受限范围扩大

2023 年 10 月 17 日，美国商务部出台了出口管制清单的 ECNN 3A090 和 4A090 要求，以进一步限制高性能 AI 芯片的出口，同时将 13 家中国公司列入实体清单。修改后的出国管制设计产品包括但不限于：英伟达 A100、A800、H100、H800、L40、L40S 以及 RTX 4090 产品。实际上，任何集成了一个或多个及以上的芯片的系统，包括但不限于英伟达 DGX、HGX 系统，都在新规涵盖范围之内。

此前，2022 年 8 月 26 日，美国政府要求英伟达停止向中国（包括中国香港）出口两款用于人工智能发展的高端计算芯片，涉及英伟达 A100 和 H100 两款芯片，以及未来推出峰值性能等同或超过 A100 的其他芯片。同时，英伟达应用这些高性能芯片的系统级产品也均在新的管制范围内。2022 年 9 月 1 日，英伟达发布声明称美国政府允许英伟达在 2023 年 9 月 1 日前，通过公司的香港工厂履行 A100 和 H100 的订单和物流运输，但售卖给中国的终端客户仍需要受美国政府批准。

表1：限制法令变化

内容	2022 年 10 月 7 日	2023 年 10 月 17 日新规
性能指标	限制每次运算位长乘以 TOPS 为单位的处理性能不能大于 4800	增加性能密度指标，同时限制总处理性能与性能密度
限制主体	拒绝英伟达向中国出口、转让的申请	扩大限制范围，不仅包括大陆与澳门特别行政区，还包括母公司设置在中国大陆以及澳门特别行政区的实体

资料来源：《对向中国出口的先进计算和半导体制造物项实施新的出口管制》(2022)，《先进计算芯片更新规则》及 《半导体制造物项更新规则》(2023)，国投证券研究中心

限制强度加大，新增多款芯片受到新规限制。根据英伟达主要芯片规格，可以计算每种芯片的性能密度指标。在新规发布之前，超过旧规性能指标限制的芯片仅为英伟达 A100，但当加入性能密度指标后，新规不仅限制了厂商出略低于性能标准的芯片以规避限制情况，同时针对数据中心芯片与非数据中心芯片进行了不同的限制约定，使更多的英伟达芯片受到禁令限制。

表2：英伟达芯片受限情况

型号	算力性能及性能密度	备注
A100	性能指标 TPP＞4800，性能密度指标＞5.92	受到新规旧规双重限制
H100	性能指标 TPP＞4800，性能密度指标＞5.92	受到新规限制
A800	性能指标 TPP＞4800，性能密度指标＞5.92	受到新规限制
H800	性能指标 TPP＞4800，性能密度指标＞5.92	受到新规限制
L40S	性能指标 TPP＜4800，性能密度处于 1.6~5.92 区间	受到新规限制
RTX4090	性能指标 TPP＜4800，性能密度处于 1.6~5.92 区间	受到新规限制

资料来源：芯东西半导体产业媒体，国投证券研究中心

1.2. 人工智能大势所趋，各地政策推进实施

“1+N”政策体系全面推动人工智能产业。2017 年国务院发布《新一代人工智能发展规划》，部委层面陆续出台相关发展规划、实施方案等落地政策，形成“1+N”政策体系，从相关法律法规和伦理规范、人工智能发展支持政策、标准和产权体系、监管和评估体系以及 AI 人才培训等五个角度全面推动人工智能健康快速发展。

同时，各一二线城市均针对 AI 产业制定了产业规模目标和企业数量目标，其中北京市于 2023 年 5 月 30 日发布《北京市加快建设具有全球影响力的人工智能创新策源地实施方案(2023-2025 年)》与《北京市促进通用人工智能创新发展的若干措施》两项重磅政策，以迅速建设具有全球广泛影响力的人工智能创新策源地。

算力发展目标明确，将带动 AI 算力的迅速发展。2023 年 10 月，工业和信息化部、中央网信办、教育部、国家卫生健康委、中国人民银行、国务院国资委等六部门联合发布《算力基础设施高质量发展行动计划》，在计算力、运载力、存储力、应用赋能等方面提出了具体目标，以进一步加强算力资源配置，提升国内算力总体水平。智算的快速发展，一方面要求智算中心的建设需要更加合理，要兼顾东西部协同发展和资源的合理利用。另一方面，智能算力更多的采用 AI 芯片，带来更大带宽的网络传输需求，这些都将显著促进 AI 芯片和网络技术的研发创新。

表3：《行动计划》概览

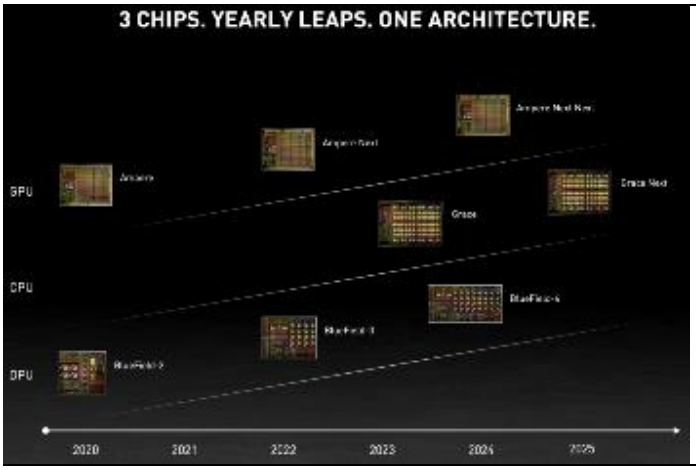
针对方面/地区		政策内容
主要目标（到 2025 年）	计算力	算力规模超过 300EFLOPS，智能算力占比达到 35%，东西部算力平衡协调发展
	运载力	数据中心集群间基本实现不高于理论时延 1.5 倍的直连网络传输，重点应用场所光传送网覆盖率达到 80%
	存储力	存储总量超过 1800EB，先进存储容量占比达到 30%以上，重点行业核心数据、重要数据覆盖率达到 100%
	应用赋能	打造一批算力新业务、新模式、新业态，工业、金融等领域算力渗透率显著提升
重点任务	京津冀、长三角、粤港澳大湾区等	面向重大区域发展战略实施需要有序建设算力设施
	贵州、内蒙古、甘肃、宁夏等节点	推进数据中心集群建设同时，着力提升算力设施利用效率，促进东西部高效互补核协同联动

资料来源：《算力基础设施高质量发展行动计划》，国投证券研究中心

2. 构建完整的芯片生态系统，“硬件+软件”缺一不可

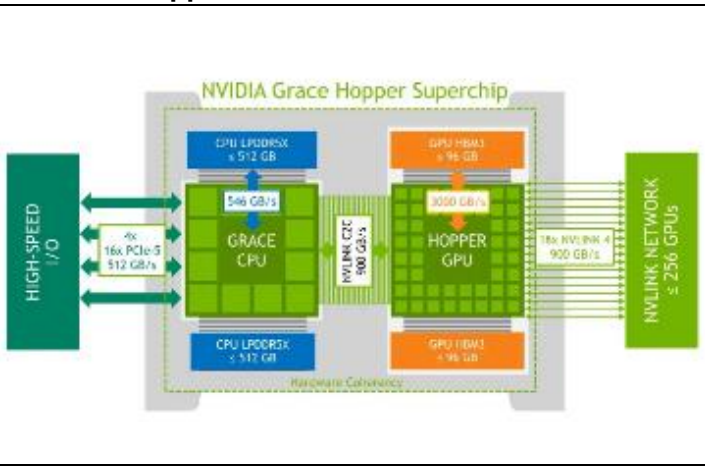
芯片方面，英伟达通过“CPU+GPU+DPU”三芯布局，数据中心正成长为公司最大业务。CPU 作为逻辑处理单元，能更好地处理各种需要快速逻辑判断和并行处理能力的计算任务；GPU 侧重对图像像素进行大规模的数据矩阵运算处理，与 AI 算法的并行结构运算匹配，GPU 在 AI 领域有着先天优势；DPU 则特别适合处理数据中心和网络设备的需求，能有效处理数据包和协议。英伟达通过将 CPU、GPU 和 DPU 集成到同一平台上，可以为客户提供更全面、高效的计算解决方案。公司推出的 Grace Hopper 超级芯片将 Grace 和 Hopper 架构相结合，为加速 AI 和高性能计算（HPC）应用提供 CPU+GPU 相结合的一致内存模型，并在大型服务器上广泛部署。2021 年及以前，游戏业务营收占比最高，但随着 AGI 引爆算力需求，公司数据中心成长极快。根据公司 2022 年年报，其数据中心业务营收约 150 亿美元（占比约 56%），已成为公司最大业务。

图1. “CPU+GPU+DPU”三芯布局



资料来源：英伟达 GTC 大会 PPT 截图、国投证券研究中心

图2. Grace Hopper 架构



资料来源：公司官网，国投证券研究中心

GPU 领域深耕数十年，长期占据市场领导地位。1999 年英伟达推出的首个 GPU 架构，开创了现代显卡的时代，代表产品是 GeForce 256 显卡，此后其架构经历了多次变革迭代，GPU 计算能力的不断提升，从 2017 到 2022 这五年间，公司先后推出了 Volta、Ampere、Hopper 等针对高性能计算和 AI 训练的架构，并以此为基础发布了 V100、A100、H100 等高端 GPU。通过不断的技术革新，英伟达 GPU 产品向量双精度浮点算力已从 7.8TFLOPS 增至 30TFLOPS。根据 Jon Peddie Research（JPR）报告显示，2023 年 Q1，英伟达 GPU 市场份额达到 84%，Q2 达到 80%，占据市场领导地位。

表4：英伟达 GPU 迭代

架构代号	Fermi	Kepler	Maxwel	Pasca	Volta	Turning	Ampere	Hopper
中文代号	费米	开普勒	麦克斯韦	帕斯卡	伏特	图灵	安培	赫柏
时间	2010	2012	2014	2016	2017	2018	2020	2022
核心参数	16 个 SM，每个 SM 包括 32 Cuda Cores 共计 512 Cuda Cores	15 个 SMx，每个 SMX 包括 192 个单精度+64 个双精度的 Cudacores ；	16 个 SMM，每个 SM 包括 4 个处理块，每个处理块包括 32 个 CUDA 内核+8 个 LD/STUnit+8 个 SFU	Pascal 架构有 GP100、GP102GP100 有 60 个 SM	80 个 SM 每个 SM 里 32 个 FP6464 个 INT3264 个 FP328 个 Tensor core	TU102 核心 72 个 SM，SM 全新设计，每个 SM 里 64 个 INT3264 个 FP328 个 Tensor core	A100 有 108 SMs 每个 SM64 个 FP3264 个 INT3232 个 FP644 个 Tensor core	H100 132 SM 每个 SM128 个 FP3264 个 INT3264 个 FP644 个 Tensor core
特点\优势	首个完整 GPU 计算架构，支持与共享存储结合纯 Cache 层次的 GPU 架构，支持 ECC 的 GPU 架构	游戏性能大幅提首次支持 GPUDirect 技术	相比 Kpler 的每组 SM 单元 192 个减少到了每组 128 个但是每个 SMM 单元拥有更多的逻辑控制电路	每个 SM 包括 64 个 cuda cores32 个 DP cores NVLink 代，双向互联带宽 160GB/sP100 有 56 个 SMHBM	Nvlink 2.0Tensor Core 1.0 满足深度学习和 AI 运算	Tensor Core 2.0RT Core 1.0	Tensor Core 3.0RT Core 2.0Nvlink 3.0 结构稀疏性 MIG 1.0	Tensor Core 4.0Nvlink 4.0 结构稀疏性知阵 MIG 2.0
纳米制程	40/28nm30 亿品体管	28nm71 亿品体管	28nm80 亿品体管	16nm153 亿晶体管	12nm211 亿晶体管	12nm186 亿晶体管	7nm283 亿品体管	4nm800 亿个晶体管
代表型号	Quadro 7000	K80，K40M	M5000M4000	P100GTX 1080P6000	V100TiTan V	T42080TIRTX 5000	A100、A303090	H100

资料来源：英伟达公司官网，国投证券研究中心

2022 年 3 月 GTC 2022 大会上，英伟达正式发布了基于 Hopper 架构的面向数据中心的新一代顶级计算核心 GH100、计算卡 H100。在机器学习及人工智能领域开放产业联盟 MLCommons 公布了最新的 MLPerf 基准评测中，英伟达 H100 Tensor Core GPU 在每次 AI 推理测试中都展现出最高性能。得益于软件优化，该 GPU 的性能比去年 9 月份首次亮相时提高了 54%，，A100 则是英伟达于 2020 年推出的上一代数据中心专用 GPU，但依然是目前 AI 训练的主流芯片产品。根据 New Street Research 的数据，英伟达占据了可用于机器学习的图形处理器市场的 95%。

表5：H100,A100,H800,A800 对比

型号	H100	A100	H800	A800
Target Markets / 目标市场	基于 GH800 图形处理器,SXM5 接口，有助于提高机器学习应用程序的速度，第四代 tensor core， 支持 FPB 和 transformer 引擎	H100 是全球范围内最大的性能出众的加速器，拥有革命性的 Transformer 引擎和高度可扩展的 NVIDIA NVLink® 互连技术等突破性功能，可推动庞大的 AI 语言模型、深度推荐系统、基因组学和复杂数字孪生的发展	新引入 FP64，TF32，BF16 Tensor Core，支持 MIG，深度学习/HPC 大规模算例并行加速计算	新引入 FP64，TF32，BF16 Tensor Core，支持 MIG，深度学习/HPC 大规模算例并行加速计算
GPU / 图形处理单元（架构）	GH100（hopper）	GH100（hopper）	GA100（Ampere）	GA100（Ampere）
GPU Cores / CUDA 单元数	16,896	16896	6,912	6,912
VRAM / 显存	80 GB	80GB	80GB	40GB
FP16 Computing（non-Tensor）/ 半精性能 non-Tensor	237.2 TFLOPS	267.6 TFLOPS	77.97 TFLOPS	70 TFLOPS

资料来源：英伟达公司官网，国投证券研究中心

NVLink 是英伟达自研的高速互连技术，解决了多 GPU 并行计算时内存共享和通信的瓶颈问题，能有效提升数据中心的整体运算能力。PCIe 是 Intel 主导的高速串行计算机扩展总线标准，是当前服务器主流的总线解决方案，PCIe 标准迭代周期约为 3 年/代，PCIe 3.0 是目前消费市场的主流选择，4.0 于 2017 年正式推出，自 2021 年下半年开始在数据中心逐步应用，并逐渐从企业级市场下沉到消费市场。目前 Intel/AMD 等主流 CPU 厂商正快速推出 PCIe 5.0 产品，用于 AI 的高性能企业级服务器通常采用 PCIe5.0 接口。NVLink 是英伟达自研的高速接口，可以提供更强大的数据传输能力和更高的吞吐量，能有效缩短数据传输时间，满足当前针对大数据和复杂运算的高带宽需求。随着 NVIDIA GPU 架构的更新和技术的不断发展，NVLink 的版本也在不断演进，以满足不断增长的计算需求和提供更优秀的性能。

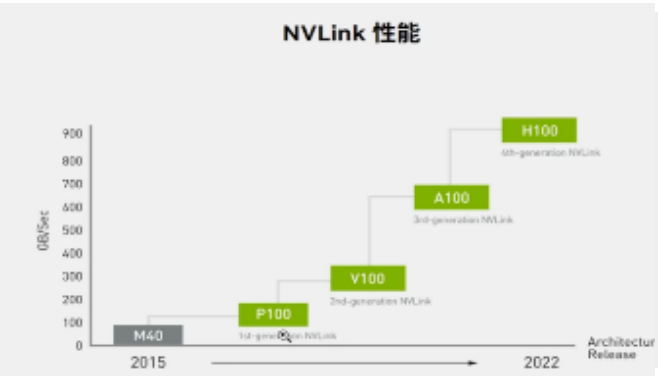
图3. NVLink 与 PCIE 带宽对比

类型	GPU	通道数	双向互联带宽
PCIE互联	A100	PCIE 4.0 x16	2GBx16x2=64GB/s
	H100	PCIE 5.0 x16	4GBx16x2=128GB/s
NVLink互联	A100	每个GPU链路Nvlink x12	25GBx12x2=600GB/s
	H100	每个GPU链路Nvlink x18	25GBx18x2=900GB/s

资料来源：英伟达公司官网，国投证券研究中心

CUDA 架构搭建英伟达软件生态，是英伟达 AI 解决方案的核心优势之一。CUDA（Compute Unified Device Architecture）是由英伟达公司推出的 GPU 加速平台，在推出 CUDA 之前，程序员想要调用 GPU 的运算资源必须先编写大量底层代码，在开发和调试上需要花费大量的时间，而 CUDA 提供了易于使用的编程模型和软件环境，允许开发者使用类似于 C/C++的高层语言进行编程，使得开发者可以借助英伟达的图形处理器（GPU）进行更为高效的并行运算。经过多年优化，目前 CUDA 已成为全球 AI 基础设施，主流的 AI 框架、库、工具都以 CUDA 为基础进行开发。国内第一批大模型厂商使用的基本都是英伟达已经构建完善的 CUDA 生态。即便竞争对手的 GPU 性能的硬件参数上比肩英伟达，如果缺乏 CUDA 的优化，其性能依然无法达到英伟达 GPU 的水平。

图4. NVLink 版本迭代



资料来源：英伟达公司官网，国投证券研究中心

图5. CUDA 架构

资料来源：英伟达公司官网，国投证券研究中心

基于其“芯片+ NVLink + CUDA”的生态系统，英伟达稳坐行业龙头地位，产品全面覆盖 AI 场景。公司 20 多年来始终引领 GPU 行业的发展，将 GPU 的主要应用场景从游戏以及画图等图像显示扩展到了以 AI、云计算等大数据相关的并行计算领域。根据 Jon Peddie Research 发布的 GPU 市场数据统计报告，英伟达 2022 年全年 PCGPU 出货量高达 3034 万块，是 AMD 的

近 4.5 倍；截至 2022 年四季度，在独立 GPU 市场，英伟达占据 84%的市场份额，远超同业竞争公司。

表6：英伟达产品概览

产品线	应用方向	产品
数据中心	AI 训练加速器 AI 推理加速器 高性能计算加速器	Volta 系列、A100 Tensor Core GPU Tesla T4、Jetson Xavier NX Tesla 系列
游戏	游戏 GPU 游戏笔记本 GPU 游戏平板 GPU	GeForce 系列 RTX 系列 Tegra K1
专业可视化	工作站 GPU 可视化集群	Quadro 系列、RTX 系列 Quadro Virtual Workstation
汽车	汽车 SoC 自动驾驶计算平台	Drive AGX 系列、Orin 系列 NVDIA DRIVE

资料来源：公司官网，国投证券研究中心

3. 华为昇腾软硬件全面布局，构建国产 AI 算力基石

3.1. 昇腾生态包括全栈的 AI 计算基础设施、行业应用及服务

昇腾生态包括昇腾系列处理器、系列硬件、CANN 异构计算架构、AI 计算框架、应用使能、开发工具链、管理运维工具、行业应用及服务等全产业链。

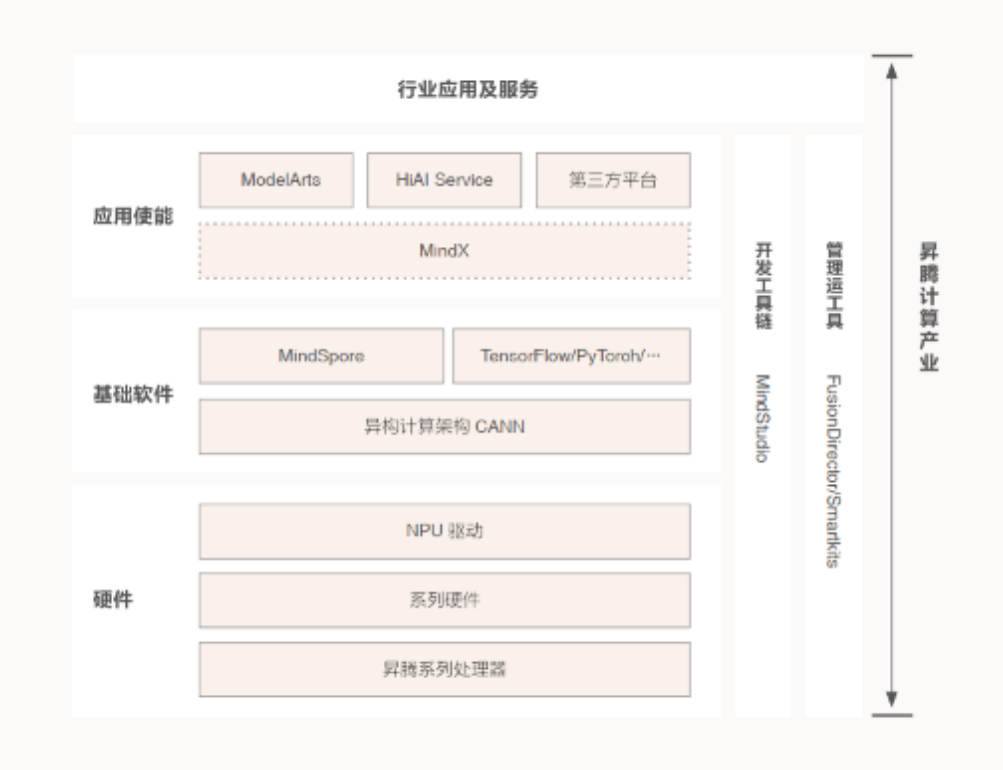
其硬件系统包括：

- 基于华为达芬奇内核的昇腾系列处理器等多样化 AI 算力；
- 给予昇腾处理器的系列硬件产品，比如嵌入式模组、板卡、小站、服务器、集群等。

其基础软件体系包括：

- 异构计算架构 CANN 以及对应的驱动、运行时、加速库、编译器、调试调优工具、开发工具链 MindStudio 和各种运维管理工具等；
- AI 计算框架，包括开源的 MindSpore，以及各种业界流行的框架，作为生态的有机组成部分。同时昇腾计算产业支持各种计算框架的对接。

图6. 昇腾生态全景



资料来源：昇腾计算产业发展白皮书，国投证券研究中心

基于达芬奇架构的昇腾芯片，运算性能优异，可应用于 AI 训练/推理场景。DaVinci 架构是面向 AI 计算设计的架构，通过独创的 16*16*16 的 3D Cube 设计，每时钟周期可以进行 4096 个 16 位半精度浮点 MAC 计算。同样是完成 4096 次运算，2D 结构需要 64 行*64 列才能计算，3D Cube 只需要 16*16*16 的结构就能算出，因此在核数与频率确定的情况下，每时钟周期进行越多的计算则算力指标越高，而 Davinci 架构的 3D 设计实现了这一点。

从性能指标上来看，昇腾 910 半精度 FP16 的运算性能可达 320TFLOPS，整型 INT8 算力可达 640TOPS，英伟达 A100 的 FP16 运算性能为 312TFLOPS。

表7：昇腾系列处理器算力详情

型号	性能参数	备注
昇腾 310	16TOPS@INT8，8TOPS@FP16	12V 主要应用在边缘计算等低功耗领域
昇腾 910	320TFLOPS@FP16，640TOPS@INT8	2019 年推出，全场景人工智能领域

资料来源：昇腾官网，国投证券研究中心

计算代价大幅缩小，功耗水平明显降低。Davinci 架构的 3D 设计以最小的计算代价增加矩阵乘的算力，实现更高的 AI 能效。2018 年 10 月华为联合奥迪展示了 L4 级无人驾驶的路测，汽车上配备了华为的 MDC 车载计算单元，但根据第五届世界互联网大会上前华为公司董事兼华为企业 BG 总裁阎力大披露，**支持 L4 级无人驾驶这样非常复杂的边缘计算场景时，昇腾 310 芯片组仅消耗共计 200 瓦的能耗**，相比英伟达系列芯片均有大幅缩减。

表8：昇腾系列与英伟达系列功耗对比

芯片	设计功耗	备注
昇腾 910	310W	
昇腾 310	8W	
昇腾 310 芯片组	200W	应用于 L4 级无人驾驶等极端复杂边缘计算场景
英伟达 H100	700W	
英伟达 A100	700W	
英伟达 H800	250W	GPU 数目较少，峰值算力低于昇腾 910
英伟达 A800	250W	峰值算力远低于昇腾 910，半精度算力仅为 280TFLOPS

资料来源：第五届世界互联网大会上前华为公司董事兼华为企业 BG 总裁阎力大发言，国投证券研究中心

HCCS 是华为自研的高速互连接口，可为内核、设备、集群提供系统内存的一致访问，片间带宽最高可达 480Gbps，是业界主流 CPU 互联速率的 2 倍多，HCCS 单个 AI 处理器提供 3 条链路能实现最多 4 个鲲鹏 920 处理器互联和最高 256 个物理核的 NUMA 架构。相比于英伟达 NVLink 与 PCIe 5.0，NVLink 单条链路双向带宽最大为 50GB/s，PCIe 5.0 仅为 4GB/s，HCCS 单条链路双向带宽可以达到 20GB/s，**HCCS 在单一链路的单向/双向互联带宽上比 PCIe 5.0 更具优势，将有效提升多个 AI 处理器协同训练的能力。**

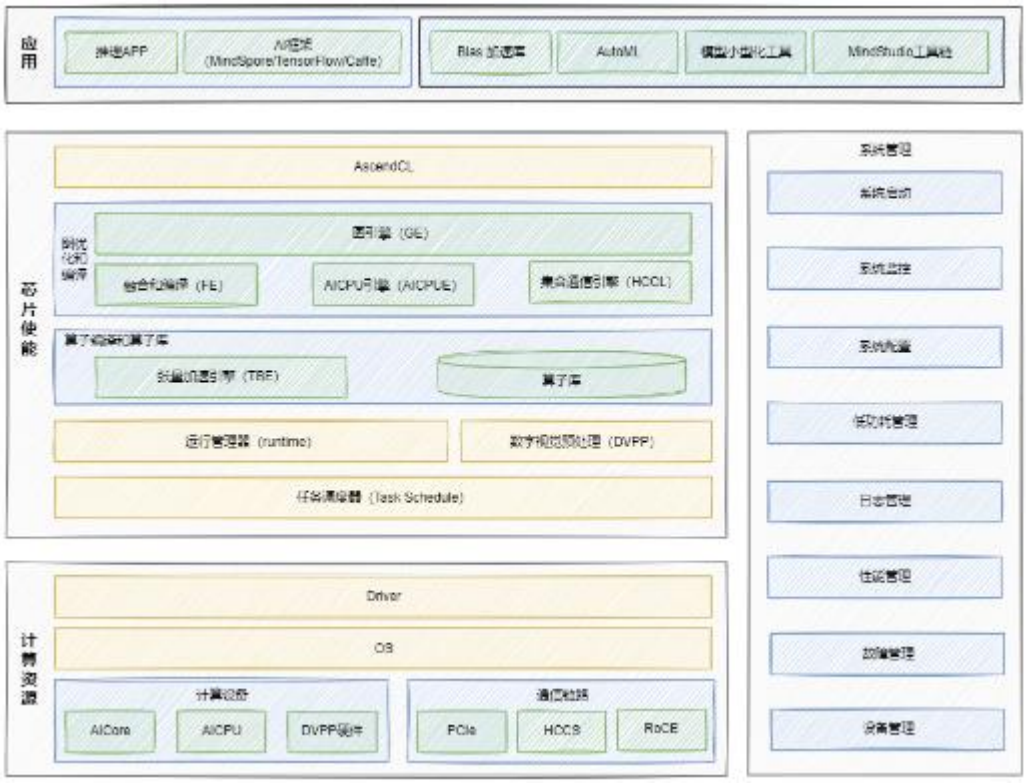
表9：HCCS 接口与 NVLink、PCIe 5.0 对比

类型	链路数	带宽
HCCS	单个 AI 处理器提供 3 条链路	每个 AI 处理器间双向互联带宽 60GB/s（3 条链路）
NVLink	最大提供 18 条链路	单链路双向互联带宽为 50GB/s,18 条链路共计 900GB/s
PCIe 5.0	最多提供 16 条链路	单链路双向互联带宽 8GB/s，16 条链路共计 128GB/s

资料来源：CSDN，NVIDIA 云数据中心，Atlas 800 训练服务器技术白皮书，国投证券研究中心

CANN 是华为针对 AI 场景推出的异构计算架构，通过提供多层次的编程接口，支持用户快速构建基于昇腾平台的 AI 应用和业务。CANN 支持端边云全场景协同，支持超过 10 种设备形态、EMUI、Andriod、openEuler、UOS、Ubuntu、Debian、Suse 等超过 14 种操作系统和多种 AI 计算框架，一套体系支持 CPU、NPU 等架构；

图7. CANN 生态



资料来源：华为云-开发者社区-昇腾 AI 基础知识介绍，国投证券研究中心

软件生态建设是华为的一大优势。为了帮助 AI 开发者更简单、更高效的开发和使用 AI 技术，华为推出面向全流程开发工具链 MindStudio。MindStudio 针对算子开发、模型训练、模型推理、应用开发、应用部署的所有全流程工具链进行整合，为开发者提供工程管理、编译、调试、运行、性能分析等全流程开发，提高开发效率。

3.2. 基于“自研芯片+自研接口+自研软件生态”，华为推出全系列解决方案

供开发者使用的高性能开发板 Atlas 200 DK，Atlas 200 DK 开发者套件（型号 3000）是以 Atlas 200 AI 加速模块（型号 3000）为核心的开发者板形态终端类产品（其中 Atlas 200 AI 加速模块是高性能 AI 计算模块，集成了昇腾 310 AI 处理器，芯片内置 2 个 AI core，可支持 128 位宽的 LPDDR4X，最大算力为 22TOPS）。

表10： Atlas 200 DK 开发者套件硬件规格

特征	规格
AI 处理器（昇腾 310AI 处理器）	2 个 DaVinci AI Core（达芬奇内核） 8 个 A55 Arm Core(最大主频 1.6GHz)
AI 算力	半精度（FP16）：4/8/11 TFLOPS 整数精度（INT8）：8/16/22 TOPS
内存	类型：LPDDR4X 位宽：128bit/64bit 容量：8GB/4GB 速率：3200Mbps 支持 ECC（Error Correction Code）
接口	千兆以太网接口：1 个 GE RJ-45 接口 USB 接口：1 个 USB3.0 Type C 接口 1 个 40pin IO 连接器，2 个 MIPI 连接器，2 个板载麦克风

资料来源：华为昇腾产品技术白皮书，国投证券研究中心

训练卡 Atlas 300 T。Atlas 300 T 训练卡（型号 9000）可以配合服务器为数据中心提供 AI 加速卡，单卡最高可提供 220 TFLOPS FP16 算力。产品具有强算力、高度集成、高速带宽等特点，可满足大量人工智能训练以及高性能计算领域的算力需求。

表11：Atlas 300 T 训练卡（型号 9000）硬件规格

特征	规格
AI 处理器（昇腾 910 AI 处理器）	30 个达芬奇 AI Core+ 16 个 TaiShan 核
内存规格	32GB HBM 16GB DDR4 2400Mbps 支持 ECC（Error Connection Code）
AI 算力	半精度（FP16）：最大 220 TFLOPS 整数精度（INT8）：最大 440 TOPS
PCIe 接口	PCIe *16 Gen4.0
网络	1*100GE QSFP-DD 接口，支持 RoCE 协议

资料来源：华为昇腾产品技术白皮书，国投证券研究中心

推理卡 Atlas 300 I 。Atlas 300I 推理卡采用 4 个昇腾 310AI 处理器的 PCIe HHHL 卡，实现快速高效的推理计算、图像识别及视频处理等工作，支持多种规格的 H.264、H.265 视频编解码。

表12：Atlas 300I 推理卡（型号 3000）硬件规格

特征	规格
AI 处理器（4 个昇腾 310AI 处理器，每个处理器包含：）	2 个 DaVinci AI Core（达芬奇内核） 8 个 A55 Arm Core(最大主频 1.6GHz)
AI 算力	半精度（FP16）：44 TFLOPS 整数精度（INT8）：88 TOPS
内存	类型：LPDDR4X 位宽：512bit 容量：32GB 速率：3200Mbps 总带宽：204.8GByte/s 支持 ECC（Error Correction Code）
PCIe 接口	PCIe3.0*8，兼容 PCIe2.0/PCIe1.0

数据来源：华为昇腾产品技术白皮书，国投证券研究中心

Atlas 800 训练服务器，8 颗昇腾算力芯片+4 颗鲲鹏 CPU。Atlas 800 训练服务器（型号：9000）是基于华为鲲鹏+昇腾处理器的 AI 训练服务器，具有超强算力密度、超高能效与高速网络带宽等特点。该服务器广泛应用于深度学习模型开发和训练，适用于智慧城市、智慧医疗、天文探索、石油勘探等需要大算力的行业领域。

表13：Atlas 800 训练服务器规格

特征	规格
CPU	4*鲲鹏 920 处理器
AI 处理器	8*昇腾处理器
内存规格	最多 32 个 DDR4 内存插槽 内存最高速率 3200MT/s 单根内存条容量支持 16/32/64GB
PCIe 拓展	最多支持 2 个 PCIe 4.0 拓展插槽
功耗	最大功耗 5.6KW

资料来源：华为昇腾产品技术白皮书，国投证券研究中心

Atlas 800 推理服务器，8 颗昇腾推理卡+2 颗鲲鹏 CPU。Atlas 800 推理服务器（型号：3000）是基于昇腾处理器的推理服务器，最大可支持 8 个 Atlas 300I 推理卡，提供强大的实时推理能力，广泛应用于中心侧 AI 推理场景。

表14：Atlas 800 推理服务器硬件规格

特征	规格
CPU	2*鲲鹏 920
AI 加速卡	最大支持 8 个 Atlas 300I 推理卡
内存规格	鲲鹏 920 7260/5250：最多 32 个 DDR4 内存插槽
PCIe 扩展槽位	鲲鹏 920 5220/3210：最多 16 个 DDR4 内存插槽 内存设计最大速率 2933MT/s 单根内存条支持 8/16/32/64/128GB 最多支持 9 个 PCIe4.0 PCIe 接口
端口	（1）前面板提供 2 个 USB 3.0 端口、1 个 DB15 VGA 端口 （2）后面板提供 2 个 USB3.0 端口，1 个 DB15 VGA 端口、1 个 RJ45 管理串口、1 个 RJ45 管理网口
显卡	显卡芯片集成在 iBMC 管理芯片中，提供 32MB 显存，支持最高 60Hz 频率下 16M 色彩的最大分辨率是 1920*1080 像素

资料来源：华为昇腾产品技术白皮书，国投证券研究中心

由数千颗昇腾处理器构成的 Atlas 900 AI 集群。Atlas 900 AI 集群由数千颗昇腾处理器构成，整合 HCCS、 PCIe 4.0 和 100G RoCE 三种高速接口。其总算力达到 256P～1024P FLOPS @FP16，相当于 50 万台 PC 的计算能力。它可以在 60 秒完成基于 Resnet-50 模型训练，比第 2 名快 15%，这可以让使用者更快的进行 AI 训练，高效地推进预测天气、勘探石油、自动驾驶等等商用进程。

图8. Atlas 900 AI 集群

资料来源：华为昇腾开发者论坛，国投证券研究中心

4. 科技巨头纷纷布局算力芯片，AI 浪潮势不可挡

4.1. AMD 的“MI 系列+Infinity Fabric+ROCm 平台”，成为英伟达全球范围内最强劲的对手

Radeon Instinct 系列是 AMD 专为数据中心和企业市场推出的 GPU 解决方案，旨在支持深度学习、高性能计算和科学研究等。从 2017 年发布 Radeon Instinct MI6，到如今更新至 Radeon Instinct MI300 系列，采用高性能的 GCN 或 RDNA 架构，支持大规模的并行计算和机器学习任务。同时支持 ROCm（Radeon Open Compute）平台，以提供开发和部署机器学习模型的工具和库。

表15：AMD Redeon Instinct 系列产品迭代

型号	发布日期	GPU 架构	光刻	显存规格	峰值功耗	计算单元	峰值性能（FP16）
MI6	06/2017	Polaris	14nm FinFET	16GB（GDDR5）	150W		5.73 TFLOPS
MI8	06/2017	GCN 3rd Gen	28nm	4GB（HBM）	175W		8.19 TFLOPS
MI25	06/2017	Vega	14nm FinFET	16GB（HBM2）	300W	64	24.6 TFLOPS
MI60	11/18/2018	Vega20	TSMC 7nm FinFET	32GB（HBM2）	300W	64	29.5 TFLOPS
MI50（16GB/32GB）	11/18/2018	Vega20	TSMC 7nm FinFET	16GB/32GB（HBM2）	300W	60	26.5 TFLOPS
MI100	11/16/2020	CDNA	TSMC 7nm FinFET	32GB（HBM2）	300W	120	184.6 TFLOPS
MI250	11/08/2021	CDNA2	TSMC 6nm FinFET	128GB（HBM2e）	560W	208	362.1 TFLOPS
MI250X	11/08/2022 1	CDNA2	TSMC 6nm FinFET	128GB（HBM2e）	560W	220	383 TFLOPS
MI210	03/22/2022	CDNA2	TSMC 6nm FinFET	64GB（HBM2e）	300W	104	181 TFLOPS
MI300X	06/13/2023	CDNA3	TSMC 5nm FinFET	192GB（HBM3）	800W	320	2615 TFLOPS

资料来源：AMD 官网，国投证券研究中心

2023 年 AMD 公司推出 Radeon Instinct MI300 系列，正式迈进“百亿亿级计算”时代。AMD Instinct MI300 系列加速器基于 AMD CDNA 3 架构打造，包括 AMD Instinct MI300A APU 加速器（创新的 AI 和 HPC 工作负载专用 APU）和 AMD Instinct MI300X GPU 加速器，可为广泛的 AI 和 HPC 工作负载提供领先的应用程序性能。随着 AI 工作负载的扩展，AMD Instinct MI300X 加速器提供了采用 UBB 业界标准 OCP 平台设计的普适性解决方案，支持客户将 8 个 GPU 整合为一个性能主导型节点，并且具有全互联式点对点环形设计，单一平台内的 HBM3 显存总计可达到 1.5 TB——提供足以应对各类 AI 或 HPC 工作负载部署的性能密集型解决方案。

2023 年 6 月，AMD 首席执行官苏姿丰（Lisa Su）在旧金山举行的发布会上表示，MI300X 提供的 HBM 密度最高是英伟达 AI 芯片 H100 的 2.4 倍，其 HBM 带宽最高是 H100 的 1.6 倍。MI300X 是针对 LLM 的优化版，拥有 192GB 的 HBM3 内存、5.2TB/秒的带宽和 896GB/秒的 Infinity Fabric 带宽。AMD 将 1530 亿个晶体管集成在共 12 个 5 纳米的小芯片中。

Infinity Fabric 是 AMD 的高速接口技术，用于连接 CPU 和 GPU 内部的不同部分，以及连接不同的 CPU 和 GPU，理论峰值 P2P I/O 带宽最高可达 896 GB/s，与 NV Link 旗鼓相当。多达 8 个 Infinity Fabric 链接将 AMD Instinct MI300X 与节点中的第三代 EPYC 处理器和其他 GPU 相连，以实现统一的 CPU 内存/GPU 显存一致性和系统吞吐量最大化，通过加速器的强大性能使 CPU 代码更简化。

表16：AMD Redeon Instinct MI300X 对比 NVIDIA H100 SXM

型号	发布日期	GPU 架构	显存规格	互联技术	接口传输速率	峰值性能（FP16）
AMD Radeon Instinct MI300X	06/13/2023	CDNA3	192GB	NV Link	896GB/s	2615 TFLOPS
NVIDIA H100 SXM	03/22/2022	Hopper	80GB	Infinity Fabric	900GB/s	1979 TFLOPS

资料来源：AMD 官网，NVIDIA 官网，国投证券研究中心

ROCm 是一个开源项目，支持多种加速器厂商和架构，提供了开放的可移植性和互操作性。作为一个开源平台，任何 CPU/GPU 供应商都可以利用 ROCm，这意味着用 CUDA 或其他平台编写的代码可以移植到供应商中立的 HIP 格式，用户可以从那里为 ROCm 平台编译代码。

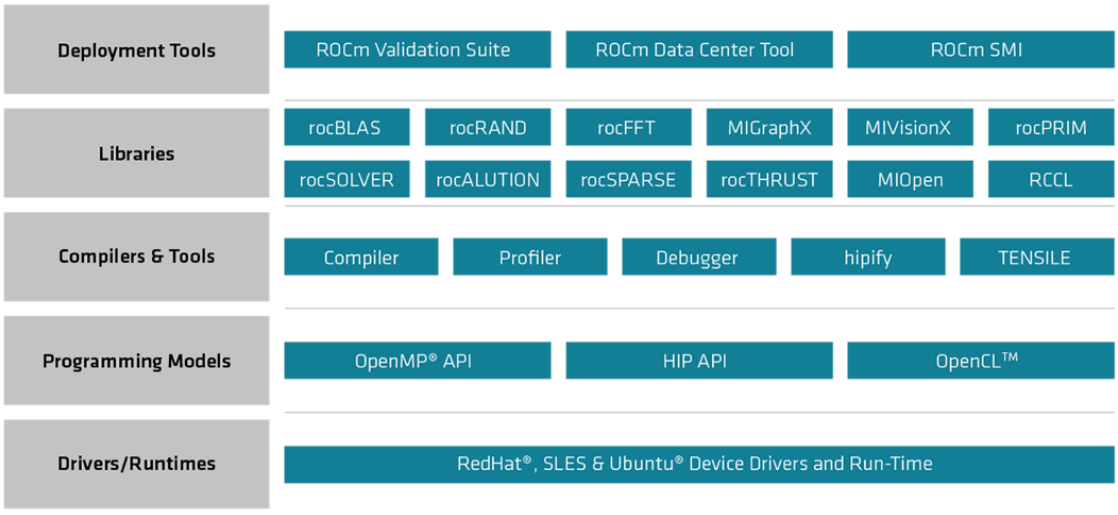
ROCm 平台针对 AMD 的 Redeon Instinct 系列有优化，并对主流机器学习框架都有所支持。AMD 将 AI 方案部署到 Radeon Instinct 系列上，使开发者可以使用 ROCm 平台在 Radeon Instinct 系列上实现更加高效和稳定的运行。同时，ROCm 平台提供了对 TensorFlow 和 PyTorch 等主要机器学习框架的原生支持，从优化的 MIOpen 库到全面的 MIVisionX 计算机视觉和机器智能库、实用程序和应用程序，AMD 与人工智能开放社区广泛合作，以促进和扩展机器和深度学习功能和优化，从而帮助扩大加速计算所适用的工作负载。

表17：AMD 相关软件生态

其他相关软件	功能
MIGraphX	图优化和编译器框架，提供高性能的模型执行
Radeon ML	机器学习库，提供针对 AMD GPU 和 CPU 的机器学习算法
Infinity Hub	云端 AI 服务平台，提供一站式的 AI 资源管理、训练、推理、部署等功能
Optimized Frameworks	提供优化后的 AI 框架，提高在 AMD GPU 上的性能和兼容性，提供预训练模型和示例代码

资料来源：AMD 官网，国投证券研究中心

图9. AMD ROCm 软件堆栈框架



资料来源：AMD 官网，国投证券研究中心

但在生态系统和性能上，ROCm 还和 CUDA 有一定差距。相比于 NVIDIA 的 CUDA，ROCm 的生态系统相对较弱，且只支持 Linux，同时更新速度较慢，生态不够完善。而在性能上，在大部分应用场景中，尽管在测试中 MI 系列的理论性能高于 NVIDIA 的加速器，但由于 ROCm 平台的优化问题，AMD 的程序性能普遍低于 NVIDIA。

4.2. 特斯拉自研 Dojo 超算服务器，芯片间高带宽互连为其一大特色

特斯拉自研宏、微架构提高运行效率与可编程性，D1 芯片在算力、互联带宽具有很强的优势。与 Nvidia，Google 等厂商的计算集群相比，Dojo 在互连、内存访问以及互联和内存访问的 IO 上是对称的，这就使得其具有独一档的 Scale Out（横向拓展）的能力，从而提高系统运行效率。微架构上，D1 芯片内部核心 Training Node 采用了图灵完备的 SMT + SIMD 设计，其可编程性有可能会强于英伟达的 Tensor Core 架构和华为的 Cube 架构，并且为每一个 Node 设计了上下左右各 64bit 的片上 NoC 通道，这使得 Node 之间核心堆叠和数据传输的难度大大降低。D1 芯片采用台积电 7nm 制程，算力达 22.6TFLOPS，总互连带宽可达 16TB/s，远超英伟达 A100 的 600GB/s、华为昇腾 910 的 90GB/s。其组成的机柜集群 Dojo ExaPOD 算力在 BF16/FP32 精度下可达到 1.1ExaFLOPs，相当于约 3200 片 A100 的算力，并拥有 1.3TB 的高速 SRAM 和 13TB 的高带宽 DRAM。同时，美国时间 2023 年 8 月 28 日，特斯拉上线了由 1 万片 H100 组成的超级计算机，将用来训练包括 FSD 自动驾驶系统在内的各种 AI 应用。目前，特斯拉的 AI 训练方向主要为自动驾驶，其硬件 Hardware 4.0 已经搭载自研 FSD 2.0。

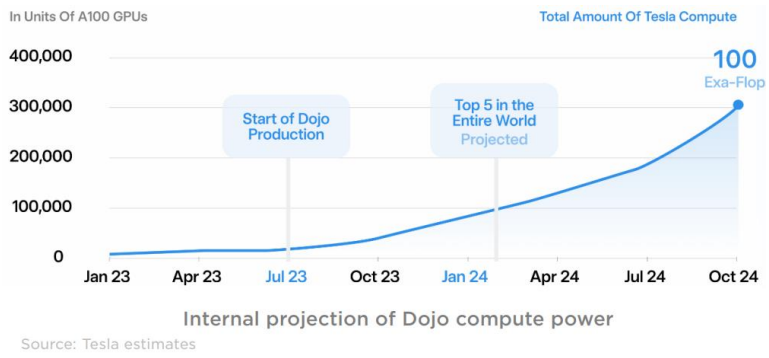
表18：Dojo ExaPOD 算力在 BF16/FP32 精度下可达 1.1ExaFLOPs

名称	算力	片上 SRAM	说明
Core	1.02TFLOPs	1.25MB	单个计算核心，Dojo 最小单位
D1	362TFLOPs	442.5MB	单芯片 D1 由 354 个实际运转 Core 构成
Tile	9.05PFLOPs	11GB	单枚训练模组 Dojo Tile 由 25 个 D1 芯片构成
Tray	54.3PFLOPs	66GB	Dojo Tray 由 6 个 Tile 构成
Cabinet	108.6PFLOPs	133GB	Dojo Cabinet 由 2 个 Tray 构成
ExaPOD	1.1EFLOPs	1.33TB	单个训练集群 Dojo ExaPOD 由 10 个 Cabinet 构成 包含 3000 个 D1，120 个训练板块

资料来源：特斯拉 AI DAY，国投证券研究中心

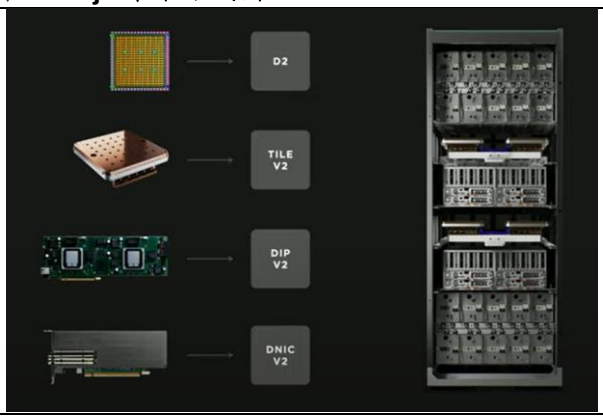
特斯拉将加快 Dojo 建设更新，全新版本 Dojo 性能将提升近十倍。据特斯拉 AI 官方账号在 Twitter 上披露的算力预期图显示，预计在 2024 年 10 月，Dojo 整体算力规模达到 100EFLOPs，相当于约 30 万块 A100 的算力总和。在 2022 年 AI Day 上，特斯拉公布了 Dojo 的未来路线规划，同时宣布公司正在研发全新版本的 Dojo 构建，包括 Dojo D2 芯片、Dojo Training Tile V2、Dip V2 和 DNIC V2。通过硬件研发更新，全新版本 Dojo 超级计算机将在性能上提升近十倍。马斯克在 Twitter 上曾表示，Dojo V1 主要面向大量的视频数据训练而优化，并不是面向通用人工智能（AGI），但 Dojo V2 将突破这一点。因此，Dojo 未来有望成为特斯拉的 AI 支柱，依靠其强大算力，特斯拉拥有的海量数据将充分释放其功能，Dojo 可全方面促进特斯拉前沿科技领域如自动驾驶、人形机器人、SpaceX 等领域。

图10. 特斯拉 Dojo 的整体算力规模将达到 100EFLOPs



资料来源：公司官网，国投证券研究中心

图11. Dojo 未来路线图



资料来源：公司官网，国投证券研究中心

自主研发 FSD2.0 芯片推动 HW4.0 更新，助力特斯拉自动驾驶突破。特斯拉在自动驾驶领域多年来不断探索，处于行业前列，其依赖于 FSD 芯片与 HW 硬件。2023 年 2 月 HW4.0 发布，其搭载了 FSD2.0 芯片。FSD2.0 芯片的 ARM Cortex-A72 CPU 内核由 12 个增加至 20 个，运行频率在 1.37GHz-2.35GHz 之间，采用 7nm 工艺，算力预测最多达到 216TOPS，仍低于英伟达 Orin254TOPS。显存方面，特斯拉成为第一个在车载领域用 GDRR 的公司。通过 FSD2.0 芯片更新升级以及 HW4.0 架构调整，特斯拉自动驾驶有望突破。

表19：Dojo ExaPOD 算力在 BF16/FP32 精度下可达 1.1ExaFLOPs

项目	HW1.0	HW2.0	HW3.0	HW4.0
发布时间	2014 年 9 月	2016 年 10 月	2019 年 4 月	2023 年 2 月
制程	40nm	16nm	14nm	7nm
核心处理器	Mobileye EyeQ3*1	Nvidia Parker SoC*1 Nvidia Pascal GPU*1 英飞凌 TriCore MCU*1	FSD1.0*2 （CPU 核*4*3） NPU*2	FSD2.0*2 （CPU 核*4*5） NPU*3
ROM	256 兆字节	6GB	LPDDR4 8*2GB	GDDR6 16*2GB
算力	0.256TOPS	21TOPS	144TOPS （双芯片）	216TOPS（预测） （双芯片）
通讯接口	N/A	N/A	1 个以太网接口	2 个以太网接口

资料来源：汽车之心，国投证券研究中心

4.3. Intel 推出 GPU 系列芯片和 oneAPI 开发平台，完善其 AI 数据中心布局

英特尔在数据中心的布局主要通过优化其至强系列 CPU 与推出数据中心 GPU max 系列和 flax 系列以及对标 DPU 的 IPU 数据接口芯片来实现。在英特尔 2022 年投资者会议上，英特尔 DCAI 公布了 2022 年至 2024 年下一代英特尔®至强产品的路线图。数据中心 GPUmax 系列有 max1550 与 max1100 两个系列，其专为 AI 和科学计算领域的数据密集型计算模型提供突破性性能。在 2021 年 6 月，英特尔首度提出 IPU 的产品概念。IPU 可以释放 CPU 的计算资源，以便于解决现代工作负载挑战，帮助提升云服务的性能。英特尔 IPU 既有基于 FPGA 的方案，如 Oak Springs Canyon，也有基于 ASIC 的方案，如 Mount Evans。

图12. 英特尔®至强产品的路线图

资料来源：公司官网，国投证券研究中心

图13. 英特尔数据中心 GPUmax 系列参数

英特尔® 数据中心 GPU Max 系列

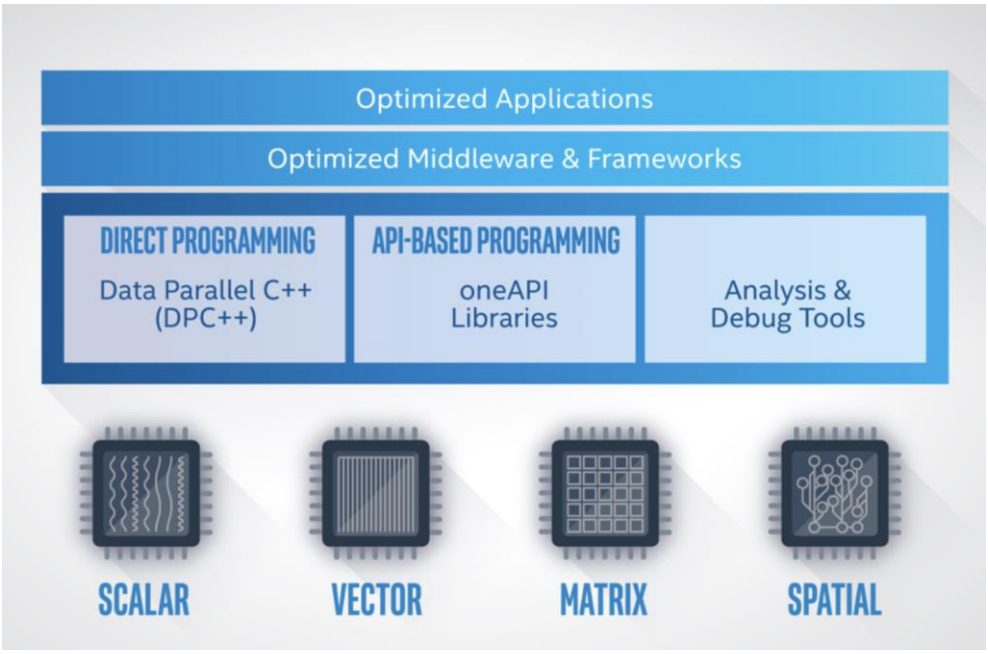
	GPU Max 1450 (600W OAM)	GPU Max 1100 (300W AIC)
架构	Xe 科学计算	
Xe 内核数	128	56
内存	HBM2E 128 GB	HBM2E 48 GB
缓存	L1 64 MB L2 408 MB	L1 28 MB L2 108 MB
最大 TDP	600W	300W
外形规格	OAM	AIC
主机互联	PCIe Gen5	
Xe Link 物理端口	Xe Link 26.5 GB/s 16 端口	Xe Link 53 GB/s 6 端口

资料来源：公司官网，国投证券研究中心

在软件生态上，英特尔推出了与英伟达的 CUDA 和 AMD 的 ROCm 对标的 oneAPI。oneAPI 是由英特尔提出的一种开放的、统一的编程模型，它旨在简化在英特尔芯片上进行并行计算的复杂性。优点是它可以跨 CPU、GPU、FPGA 和其他硬件架构运行。

通过这种方式，一套代码就可以应用于多种硬件平台，实现跨平台的并行计算，大大提高了开发效率和应用性能，其精简程度对标英伟达 CUDA 架构与 AMD 的 ROCm。

图14. OneAPI 架构

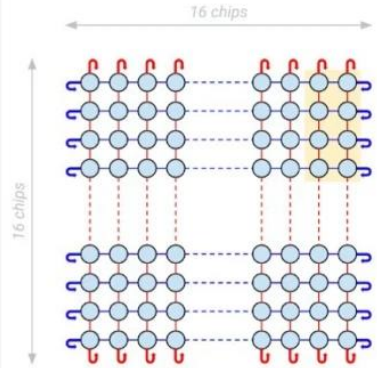


资料来源：英特尔 OneAPI 规范 1.3 版本原文，国投证券研究中心

4.4. Google 推出 Cloud TPU 解决方案，更专注于机器学习领域

Google Cloud TPU v5e 专为提供大中型训练与推理所需的成本效益和性能而设计。性能上，TPU v5e 能够平衡性能、灵活性和效率，允许多达 256 个芯片互连，聚合带宽超过 400 Tb/s 和 100 petaOps 的 INT8 性能，更加高的带宽，对于大模型的数据传输非常有利。Cloud TPU v5e 通过基准测试显示，大模型的训练速度提高了 5 倍。在推理指标方面，Cloud TPU v5e 也实现了巨大的提升，能够每秒实时处理 1000 秒的内部语音到文本和情感预测模型。

图15. TPU v5e 芯片互联效率提升



TPU v5e Chips	Max Model Size* (# parameters)	Reference Models & Sizes
1	13 Billion	LLaMA 2 13B
4	32.5 Billion	LLaMA 32.5B
8	65 Billion	LLaMA 65B
16	175 Billion	GPT-3 175B
32	280 Billion	Gopher 280B
64	540 Billion	PaLM 540B
128	1 Trillion	GLaM 1T
256 (1 Pod)	2 Trillion	Switch Transformer 1.6T

资料来源：谷歌数据中心，国投证券研究中心

和英伟达通用型 GPU 相比，谷歌 TPU 更加专注深度学习领域，可加快深度学习运算速度，降低功耗。英伟达的 A100 和 H100 属于广义上的通用型 GPU，GPU 具有通用计算能力，适用于高性能计算、深度学习以及大规模数据分析等多种计算工作负载，而不仅仅是深度学习。谷歌 TPU 采用低精度计算，在几乎不影响深度学习处理效果的前提下大幅降低了功耗、加快运算速度，尤其对于中型 LLM 设计者来说完全够用，因此他们可能不需要依赖高性能的英伟达 A100/H100。同时，TPU 使用了脉动阵列等设计来优化矩阵乘法与卷积运算。

表20：Google TPU v5e 对比 NVIDIA H100 SXM

型号	发布日期	显存规格	峰值性能（FP16）
Google TPU v5e	2023	95GB	459 TFLOPS
NVIDIA H100 SXM	2022	80GB	1979 TFLOPS

资料来源：谷歌数据中心，NVIDIA 官网，国投证券研究中心

TensorFlow 框架的广泛性、泛用性为 Google 构筑 AI 护城河。TensorFlow 是由 Google 团队开发的最重要的深度学习框架，也是全世界使用人数最多、社区最为庞大的一个框架。Tensorflow 灵活的架构可以部署在一个或多个 CPU、GPU 的台式及服务器中，支持多家 NVIDIA、AMD 等多家厂商的 GPU 加速器，或者使用单一的 API 应用在移动设备中。

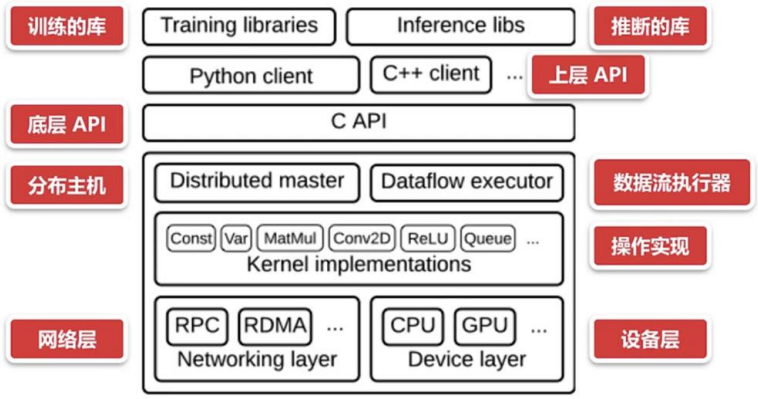
表21：Google TPU v5e 对比 NVIDIA H100 SXM

其他相关软件	功能
Cloud AI Platform	一个云端的 AI 服务平台，提供了一系列的工具和服务，包括数据标注、数据集管理、模型管理、自动化机器学习、在线预测、批量预测等
BigQuery ML	一种在 BigQuery 中直接使用 SQL 语句进行机器学习的功能，可以利用 BigQuery 的强大的数据分析能力，快速创建和部署机器学习模型
AI Hub	一个在线的 AI 资源库，可以分享和发现各种 AI 相关的资产，如数据集、模型、管道、笔记本等
Kubeflow	一个基于 Kubernetes 的开源平台，可以轻松地构建可移植和可扩展的机器学习工作流

资料来源：谷歌数据中心，TensorFlow 官网，国投证券研究中心

Google 以软件优势带动硬件发展，以 TensorFlow 框架助力 Cloud TPU。Google 的 Cloud TPU 系列加速器经过优化，可加速和扩展使用 TensorFlow 编程的特定 ML 工作负载。Cloud TPU 还简化了对 ML 计算资源的计算和管理，可使得 ML 模型加速最优化，并根据需求动态调整容量；Cloud TPU 的大规模、高集群的 ML 模型已经过多年优化，无需投入专门的能源、冷却、网络和存储设备等方面的精力、时间和专业知识来进行设计、安装和维护。

图16. TensorFlow 详细架构



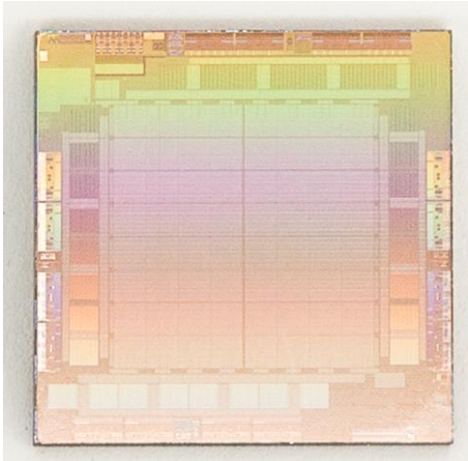
资料来源：alanhou 网，国投证券研究中心

4.5. Meta 2020 年推出第一代 MTIA 方案，侧重于处理低/中复杂度模型

作为 Facebook 母公司，META 在 AR/VR 头显全球市场上有明显份额优势，并开发有 Llama 2、Falcon 40B、Stable Diffusion 等 LLM 模型；为实现差异化竞争、提高公司核心实力，META 在 AI、AR 芯片生态领域持续探索自研，目前已推出 AI 推理定制芯片 MTIA v1、RSC 超算集群与深度学习框架 Pytorch、Caffe2go 等。

2020 年 Meta 推出第一代 MTIA 全栈解决方案，包括 MTIA 芯片、PyTorch 与推荐模型，目前侧重于处理低/中复杂度的 AI 模型。其中，加速器采用台积电 7nm 工艺制造，具备 800 MHz 的运行功率，在 INT8、FP16 精度下分别能够提供 102.4 TOPS、51.2 TFLOPS 算力。MTIA 第一代致力于提高推荐模型效率、应用于广告及其他新闻推送，采用开源芯片架构 RISC-V，功耗仅有 25 瓦，远低于英伟达等主流厂商的芯片产品；在基准测试中第一代 MTIA 芯片**处理低/中等复杂度的 AI 模型效率高于 GPU**，在这方面相较竞品芯片有明显优势。

图17. MITA V1



资料来源：Meta 官网，国投证券研究中心

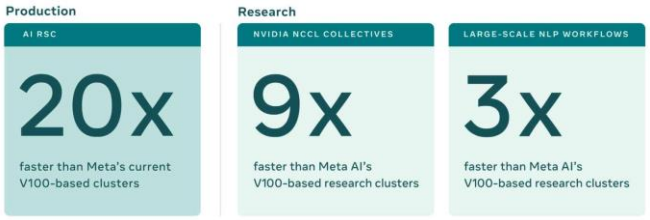
图18. MTIA 的深度学习推荐模型（DLRM）端到端性能结果

DLRM model	Size (GB)	Complexity (GFLOPs/batch)
Low complexity 1	53.2	0.032
Low complexity 2	4.5	0.014
Medium complexity 1	120	0.140
Medium complexity 2	200	0.220
High complexity	725	0.450

资料来源：Meta 官网，国投证券研究中心

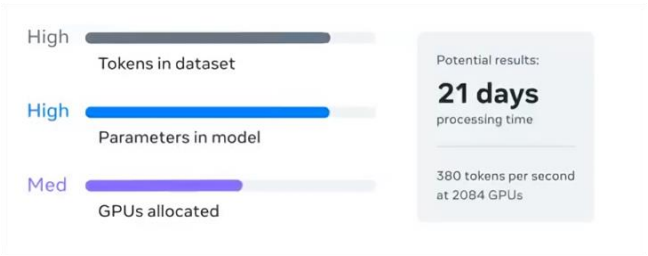
根据 Facebook 官网披露，RSC AI 服务器系列由 META 与 Penguin Computing、英伟达、Pure Storage 合作组装，于 2022 年 1 月首次亮相。目前，RSC 包含 2000 个英伟达 DGX A100 系统和 16000 个英伟达 A100 GPU，能实现近 5 exaflops 的混合精度算力，已用于推动包括生成式 AI 在内的多领域前沿研究。Meta 将 RSC 应用于训练有 650 亿参数的大语言模型 LLaMA 并将该模型作为门控版本分享给研究社区，以帮助研究人员在无大量硬件的情况下对特定任务进行研究、微调。

图19. AI RSC 与基于 V100 的集群的运算速度对比



资料来源：Meta 官网，国投证券研究中心

图20. RSC 计算性能



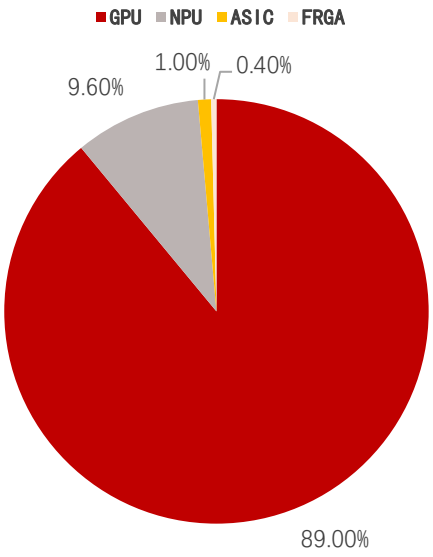
资料来源：Meta 官网，国投证券研究中心

5. AI 产业带动国内算力数据中心建设，大规模招标陆续启动

AI 产业崛起，国产 AI 芯片和服务器的需求在快速增长。2023 年 11 月 29 日，在 AICC 2023 中国人工智能算力大会上，国际数据公司（IDC）与浪潮信息联合发布《2023-2024 中国人工智能计算力发展评估报告》。根据报告，2023 年中国 AI 服务器市场规模将达 91 亿美元，同比增长 82.5%，智能算力规模预计达到 414.1EFLOPS，同比增长 59.3%。在 GPU、CPU 领域，国产厂商如华为、中科曙光、兆芯、海光等纷纷推出自主研发的芯片产品，打破了国外厂商在高端芯片市场的垄断。

在算力需求不断增长的大背景下，数据中心的建设也在加速进行。新技术和新应用的快速发展，如工业互联网、云计算、大数据等，加上 ChatGPT 技术的推广，对数据资源的存储、计算和应用需求提出了更高要求。国内外的数据中心建设有望迎来一个高峰期。根据《新型数据中心发展三年行动计划（2021-2023 年）》、《数字中国发展报告（2022 年）》以及共研产业咨询数据，到 2023 年，我国数据中心机架行业市场规模将达到 750 万架，市场规模预计将达到 2470.1 亿元人民币。

图21. 2022 年中国人工智能芯片规模占比



资料来源：IDC、国投证券研究中心

根据各大科技巨头公布的自研芯片性能参数，我们把运算性能用半精度 FP16 或 FP32 来统一比较，得到表格如下：

表22：各厂商芯片性能对比

产品	推出时间	半精度 FP16/TFLOPS
昇腾 910	2019 年 8 月 23 日	360 TFLOPS
V100	2017 年 5 月 11 日	28 TFLOPS
A100	2020 年 5 月	312 TFLOPS
H100	2022 年 3 月 22 日	1979 TFLOPS
H200	2023 年 11 月 13 日	1979 TFLOPS
思元 590	2023 年 4 月 18 日	＞312 TFLOPS
壁仞 BR100	2023 年 8 月 9 日	256 FP32 TFLOPS
海光 DCU8100	---	≈70%*312 TFLOPS
沐曦 MXC500	2023 年 6 月 13 日	FP32 15TFLOPS
Dojo D1	2021 年 8 月 19 日	FP32 22.6 TFLOPS
AMD MI300X	2023 年 12 月 7 日	FP32 45.3 RFLOPS

资料来源：各公司官网，国投证券研究中

2023 年 10 月 15 日，中国电信官网披露，AI 算力服务器（2023-2024 年）集中采购项目总计 4157 台，预计采购总额超过 80 亿元，根据评审结果，超聚变、浪潮、新华三等厂商入围，并显著增加了对训练型服务器的投资。这一采购规模与 2021-2022 年仅采购 1268 台 GPU 型服务器相比，有显著增长。

图22. 2023-2024 中国电信 AI 算力服务器集采中标候选人表

中国电信 AI 算力服务器（2023—2024 年）集采中标候选人			
标包	排名	企业名称	投标限价（元，含增值税）
训练型风冷服务器（I 系列）	1	超聚变数字技术有限公司	5,342,130,820.87
	2	浪潮电子信息产业股份有限公司	5,376,127,950.02
	3	紫光华山科技有限公司	5,365,504,587.24
	4	宁畅信息产业（北京）有限公司	5,384,213,138.44
	5	中兴通讯股份有限公司	5,285,535,434.18
	6	烽火通信科技股份有限公司	5,040,734,142.75
	7	联想（北京）信息技术有限公司	5,208,044,679.61
训练型液冷服务器（I 系列）	1	超聚变数字技术有限公司	341,995,767.86
	2	浪潮电子信息产业股份有限公司	344,327,286.69
	3	紫光华山科技有限公司	342,200,460.58
	4	宁畅信息产业（北京）有限公司	342,729,177.41
训练型风冷服务器（G 系列）	1	四川华鲲振宇智能科技有限责任公司	1,304,993,691.62
	2	河南昆仑技术有限公司	1,301,333,577.55
	3	烽火通信科技股份有限公司	1,301,966,880.40
	4	宝德计算机系统股份有限公司	1,308,854,254.13
	5	新华三信息技术有限公司	1,301,879,376.59
	6	湖南湘江鲲鹏信息科技有限责任公司	1,300,296,513.27
	7	北京神州数码云科信息技术有限公司	1,301,661,686.61
	8	黄河科技集团信息产业发展有限公司	1,307,306,069.38
训练型液冷服务器（G 系列）	1	四川华鲲振宇智能科技有限责任公司	1,477,099,321.33
	2	河南昆仑技术有限公司	1,475,413,614.45
	3	烽火通信科技股份有限公司	1,475,489,300.72
	4	新华三信息技术有限公司	1,475,901,834.34
	5	宝德计算机系统股份有限公司	1,484,725,722.97
	6	湖南湘江鲲鹏信息科技有限责任公司	1,474,476,346.12
	7	北京神州数码云科信息技术有限公司	1,476,210,004.55
	8	黄河科技集团信息产业发展有限公司	1,484,431,732.00

资料来源：中国电信官网、新浪科技、国投证券研究中心

具体到各标包的分布，标包 1 和标包 2 均属于 I 系列服务器，包括 2073 台训练型风冷服务器、125 台训练型液冷服务器和 1182 台 InfiniBand 交换机。而标包 3 和标包 4 则为 G 系列服务器，分别包括 1048 台训练型风冷服务器和 929 台训练型液冷服务器。

中标厂商主要为国内服务器集成商，为国产芯片导入营造良好的环境。超聚变、浪潮信息、紫光华山、宁畅、中兴通讯、烽火通信和联想在 I 系列风冷和液冷服务器的投标中各占一席之地，超聚变在风冷服务器和液冷服务器的中标金额和市场份额中均是第一。另一方面，在 G 系列服务器的标包中，华鲲振宇、昆仑、烽火通信、宝德计算、新华三、湘江鲲鹏、DCN 和黄河信产等公司均成为中标候选人。

表23：中国电信此次 AI 服务器集 4 个标包情况

标包类别	企业名称	报价/中标金额（亿元）	份额
标包 1：训练型风冷服务器(I 系列)	超聚变	16.56	31%
	浪潮信息	11.75	22%
	紫光华山	8.6	16%
	宁畅	6.44	12%
	中兴通讯	5.38	10%
	烽火通信	2.64	5%
	联想	2.02	4%
标包 2：训练型液冷服务器(I 系列)	超聚变	1.37	40%
	浪潮信息	1.03	30%
	紫光华山	0.68	20%
	宁畅	0.34	10%
标包 3：训练型风冷服务器(G 系列)	华鲲振宇	13.05	-
	昆仑	13.01	-
	烽火通信	13.02	-
	宝德计算	13.09	-
	新华三	13.02	-
	湘江鲲鹏	13	-
	DCN	13.02	-
	黄河信产	13.07	-
标包 4：训练型液冷服务器(G 系列)	华鲲振宇	14.77	-
	昆仑	14.75	-
	烽火通信	14.75	-
	新华三	14.76	-
	宝德计算	14.85	-
	湘江鲲鹏	14.74	-
	DCN	14.76	-
	黄河信产	14.84	-

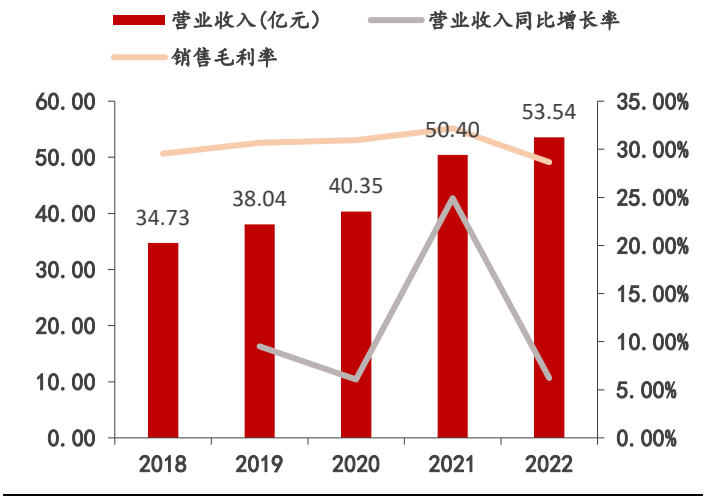
资料来源：中国电信公告，国投证券研究中心

6. 相关公司

6.1. 兴森科技

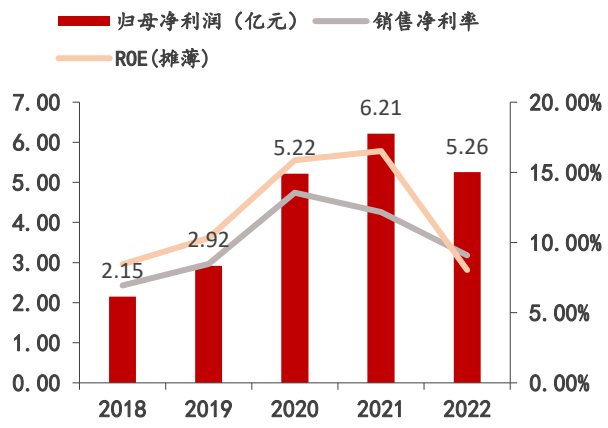
兴森科技是国内领先的印制电路板（PCB）样板及批量板的设计和制造服务提供商，公司于2012 年开始涉足 CSP 封装基板领域，是国内 IC 封装基板行业的先行者之一。目前，公司在薄板加工能力和精细路线能力方面居于国内领先地位，并与国内外主流的芯片厂商、封装厂建立了合作关系，积极投入 FCBGA 封装基板领域。

图23. 兴森科技营收及同比增速



资料来源：Wind、国投证券研究中心

图24. 兴森科技归母净利润及同比增速

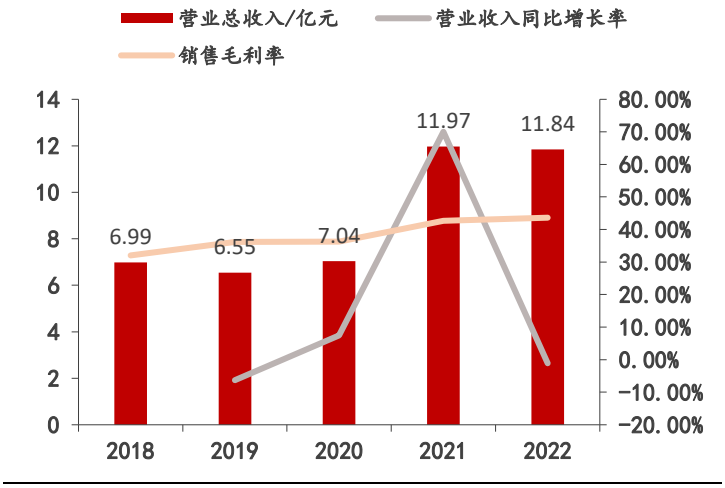


资料来源：Wind、国投证券研究中心

6.2. 新益昌

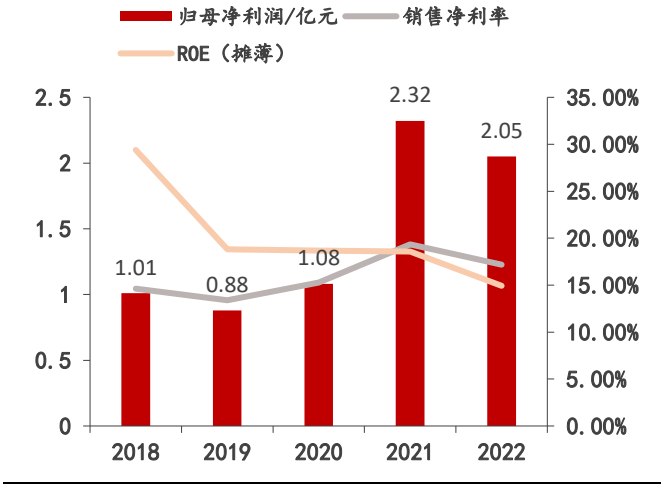
新益昌是国内领先的 LED 和半导体固晶机综合解决方案提供商，随近年来 3C 电子产品需求增加，同时以云计算、大数据、新能源及可穿戴设备等为主的新型应用领域强劲需求下，半导体市场出现巨大需求。根据 WSTS 预测，2024 年全球半导体销售额预计可回升至 5760 亿美元。其中，半导体封装环节的重点是固晶及焊线环节，固晶环节对设备的超高精度、定位能力具有极高的要求，技术壁垒很高，是公司的核心竞争力所在。根据公司 2023 半年报披露，公司凭借过硬的产品质量和技术创新能力以及配套服务能力，在半导体封装领域为晶导微、灿瑞科技、扬杰科技、通富微、固锝电子、华天科技等知名公司在内的庞大优质客户群体提供定制化服务。目前，公司半导体固晶设备近年来客户导入顺利，受到业内认可，业务收入快速增长，根据公司 12 月 27 日发布的自愿披露订单情况，公司截至 12 月 26 日，固晶机板块在手订单共计 4.13 亿元。

图25. 新益昌营收及同比增速



资料来源：Wind、国投证券研究中心

图26. 新益昌归母净利润及同比增速

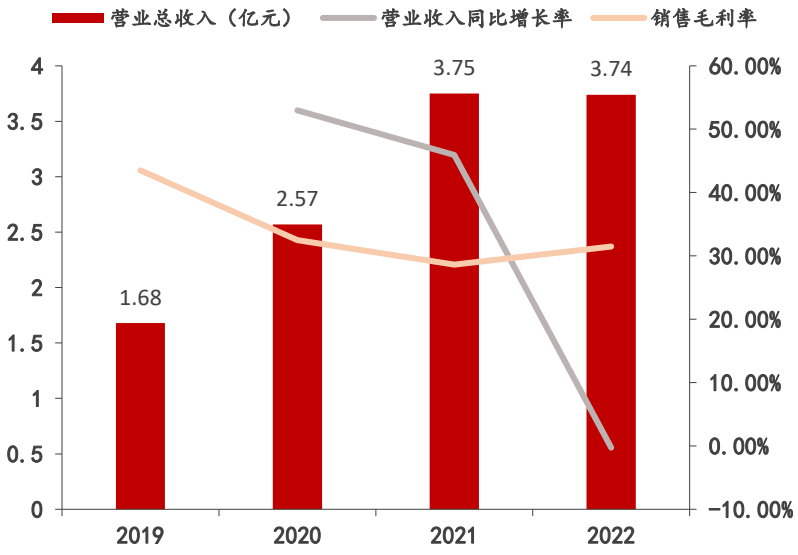


资料来源：Wind、国投证券研究中心

6.3. 天承科技

天承科技成立于 2010 年，主要经营 PCB 所需要的专用电子化学品的研发、生产和销售（功能湿电子化学品）。PCB 专用电子化学品专用性强、品种多，公司经过多年积累，目前技术已经涵盖水平沉铜、电镀、垂直沉铜、化学沉锡、去膜、棕化、粗化、微蚀等多个 PCB 制作工艺流程，技术指标与应用性能达到行业先进水平，能够满足下游厂商对于生产高频高速 PCB、HDI、多层软板及软硬结合板等高端 PCB 需求，同时公司也解决了触摸屏金属网格沉铜，品牌知名度较高。

图27. 天承科技营收及同比增长



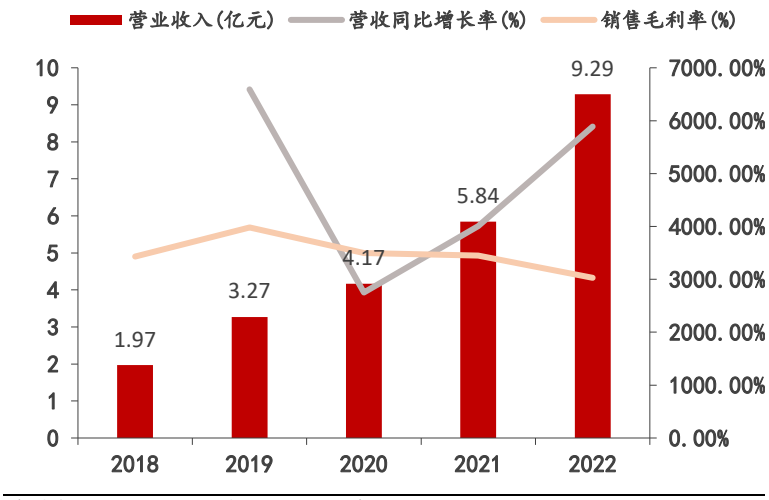
资料来源：Wind，国投证券研究中心

6.4. 德邦科技

公司以其在高端电子封装材料研发及产业化领域的专业实力，已被认定为国家级的专精特新“小巨人”企业，其业务聚焦于集成电路封装材料、智能终端封装材料、新能源应用材料及高端装备应用材料等四大类别。这些产品广泛用于不同的封装工艺环节和应用场景，如晶圆加工、芯片级封装、功率器件封装、板级封装、模组及系统集成封装等，在国家集成电路产

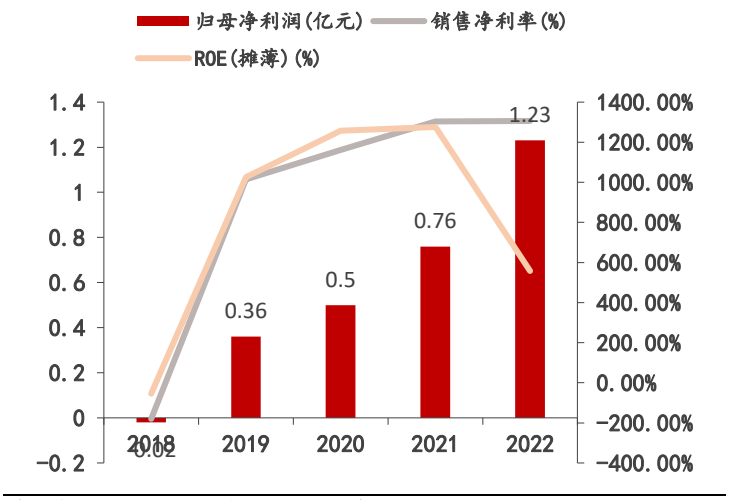
业基金的重点支持下，已在集成电路封装、智能终端封装、动力电池封装、光伏叠瓦封装等领域取得显著的技术突破。

图28. 德邦科技营收及同比增速（单位:亿元）



资料来源：Wind，公司年报，国投证券研究中心

图29. 德邦科技归母净利润及同比增速（单位:亿元）

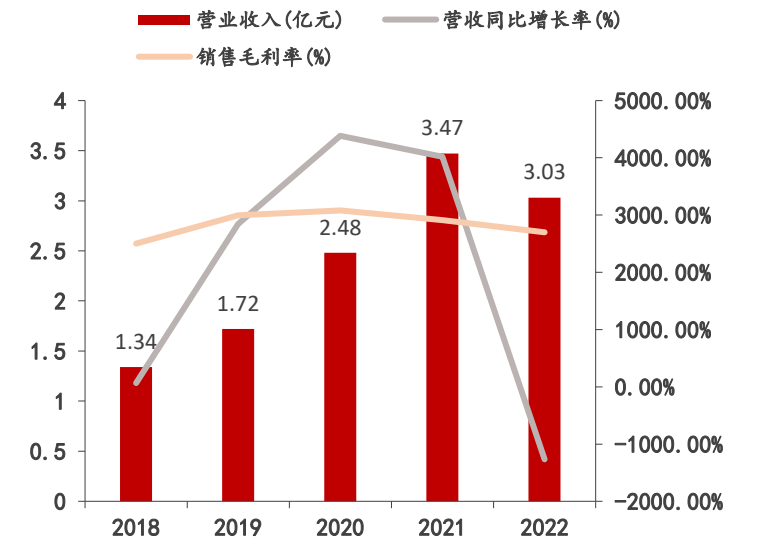


资料来源：Wind，公司年报，国投证券研究中心

6.5. 华海诚科

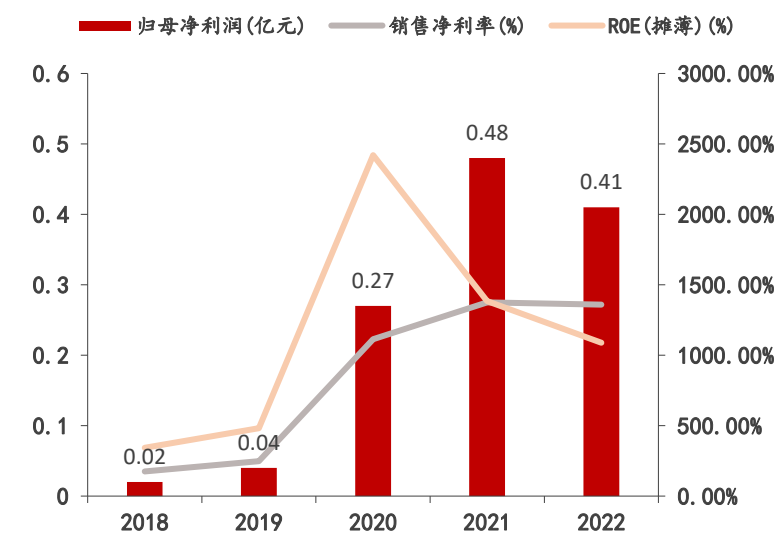
公司是一家专注于半导体封装材料的研发及产业化的国家级专精特新“小巨人”企业，以其在环氧塑封料领域的显著成就而闻名。在先进封装领域，华海诚科也取得了显著进展。据公司业绩会披露，其应用于 QN 封装的产品已通过通富微电和长电科技等知名企业的验收，并开始小批量销售。此外，跟据公司招股书,FC 底填胶等多款产品也已进入小批量生产和销售阶段，而应用于 FCBGA 的产品和液态塑封材料(LMC)目前正处于客户验证阶段，2021 年，公司成为长电科技、华天科技等多家知名封装企业的主要内资供应商。

图30. 华海诚科营收及同比增速（单位:亿元）



资料来源：Wind，公司年报，国投证券研究中心

图31. 华海诚科归母净利润及同比增速（单位:亿元）



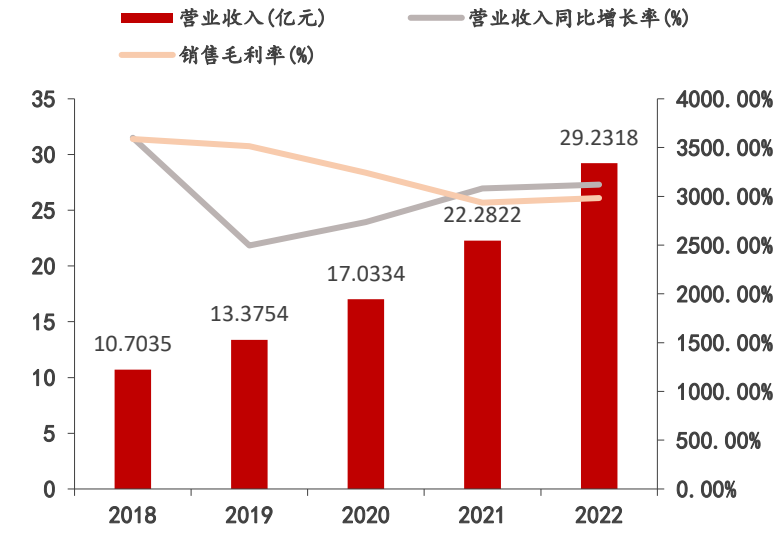
资料来源：Wind，公司年报，国投证券研究中心

6.6. 英维克

液冷系统是一种先进的散热技术，它通过利用液体的高导热系数，有效提升了散热效率，尤其在处理高功率、高热量的 AI 服务器和 GPU/CPU 等高性能计算设备时表现卓越。

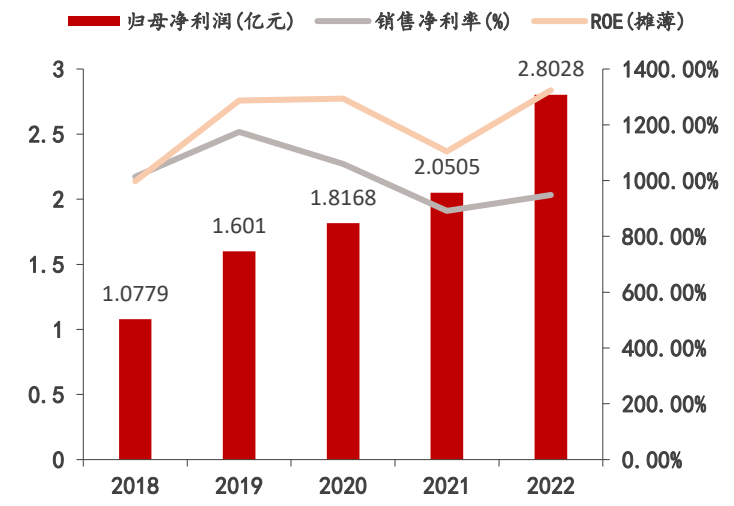
2005 年创立以来，深圳英维克公司在精密温控节能设备领域持续发展和创新，成为该领域在国内的技术领导者。最初，该公司专注于为信息和通信行业提供温控解决方案，产品范围涵盖了通信基站、户外机柜和数据中心节能空调等，根据公司 2020 年报披露，公司与华为、中兴等行业巨头建立了深入合作关系。2013 年，英维克洞察市场前景，进入储能温控领域，到 2020 年，已牢固占据行业领先地位。2015 年，公司进一步扩张，设立深圳科泰，进军新能源车空调领域，并于 2018 年收购上海科泰，拓展至轨道交通领域。在不断拓宽温控解决方案的同时，公司通过构建平台化解决方案，发挥了协同效应，增强了市场竞争力。

图32. 英维克营收及同比增速 (单位:亿元)



资料来源：Wind，公司年报，国投证券研究中心

图33. 英维克归母净利润及同比增速 (单位:亿元)



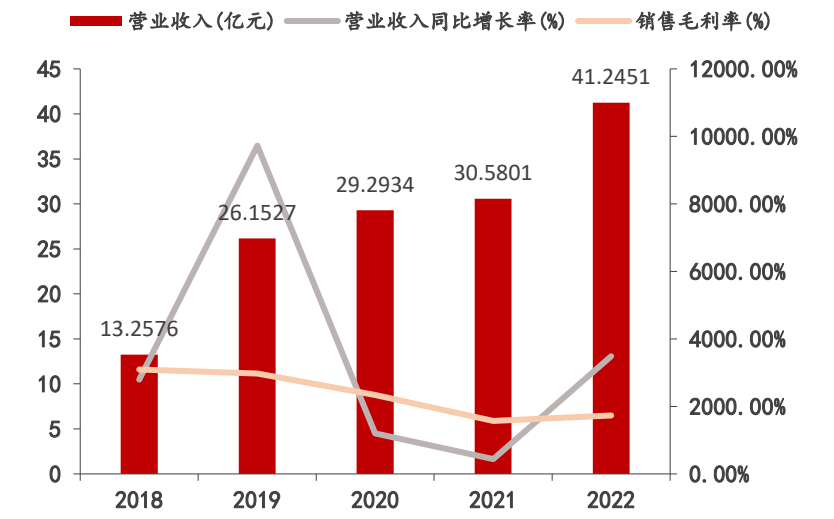
资料来源：Wind，公司年报，国投证券研究中心

6.7. 飞荣达

自 1993 年成立以来，公司在电子辅料产品生产领域取得了显著发展。自 2000 年开始，公司着手开发国际主流品牌的电磁屏蔽和导热材料与相关散热解决方案，逐步扩展其产品线至金属屏蔽器件、导电布衬垫、导热界面器件等，拓展应用至计算机和手机等高端领域。据公司公告与 2022 年年报，在通信基站领域，通过与中兴通讯的合作，公司成功开发并试制了 3D VC 技术样机，标志着 5G 基站首次采用这一先进的散热技术。在服务器领域，飞荣达向包括华为和超聚变在内的多个知名客户提供了一系列散热与电磁屏蔽解决方案和产品。其产品线涵盖单相和两相液冷模组、轴流风扇以及专门设计的散热器等，以满足客户多样化的散热需求。

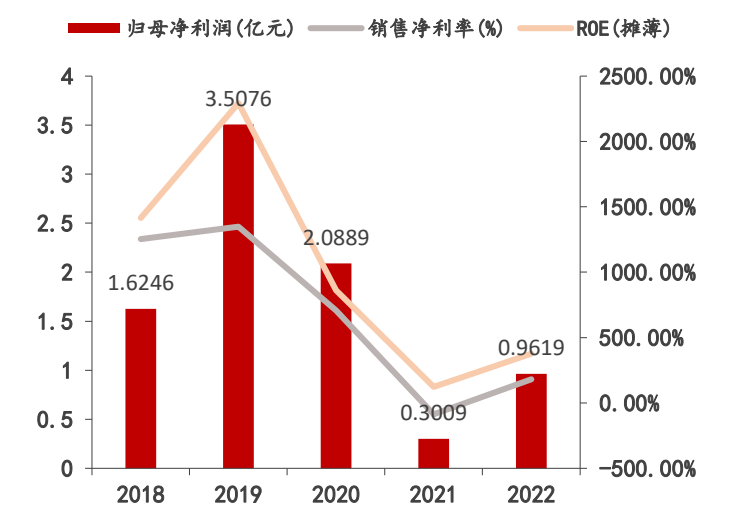
公司不仅拥有华为、中兴、微软等国内外大型企业客户，还成功扩展至 Facebook 和 Google 等全球知名企业。这一广泛的客户基础证明了其产品的实力和市场认可度。飞荣达的优势还体现在其完整的散热产业链布局上，公司不仅在散热器件和材料领域具有强劲竞争力，而且上下游一体化的产业链布局进一步增强了公司在整个散热行业的综合竞争优势。随着液冷散热市场的不断扩大及市场潜力逐步释放，飞荣达有望获得显著的市场机遇。

图34. 飞荣达营收及同比增速（单位:亿元）



资料来源：Wind，公司年报，国投证券研究中心

图35. 飞荣达归母净利润及同比增速（单位:亿元）



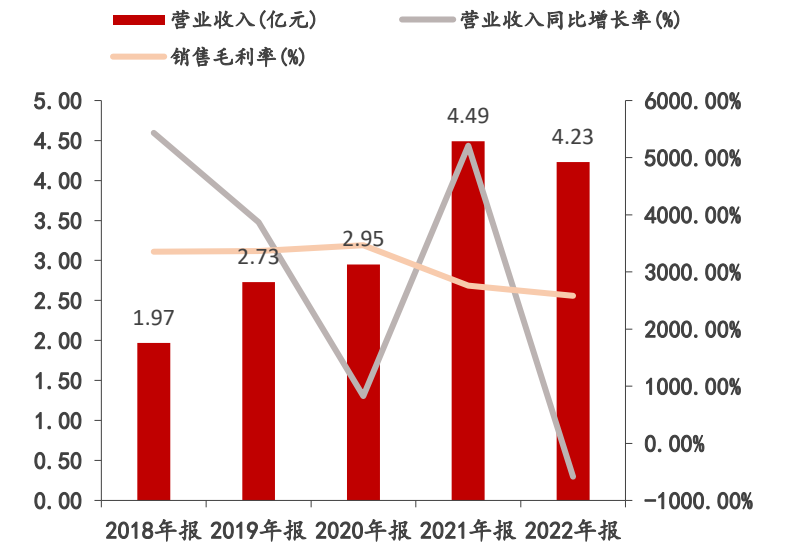
资料来源：Wind，公司年报，国投证券研究中心

6.8. 思泉新材

思泉新材是一家专注于热管理材料的多元化功能性材料提供商，在电子电气产品稳定性及可靠性提升方面表现卓越。其主营业务包括研发、生产和销售热管理材料、磁性材料、纳米防护材料等。作为国内电子电气功能性材料领域的领先高新技术企业，思泉新材在自主研发和技术创新方面具有显著优势。

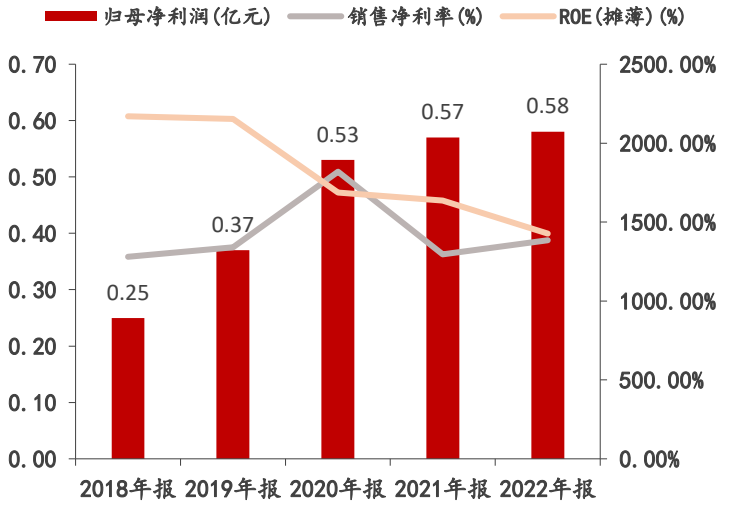
公司在行业内获得了多项殊荣和认可，包括“专精特新‘小巨人’企业”称号和“2021 年广东省专精特新中小企业”。此外，思泉新材还拥有“广东省石墨散热复合材料工程技术研究中心”和“广东省博士工作站”，并被认定为“2020 年度广东省知识产权示范企业”。至 2022 年底，公司共拥有 73 项专利，其中包括 22 项发明专利，展现了其在知识产权保护方面的重视和成效。

图36. 思泉新材营收及同比增速（单位:亿元）



资料来源：Wind，公司年报，国投证券研究中心

图37. 思泉新材归母净利润及同比增速（单位:亿元）

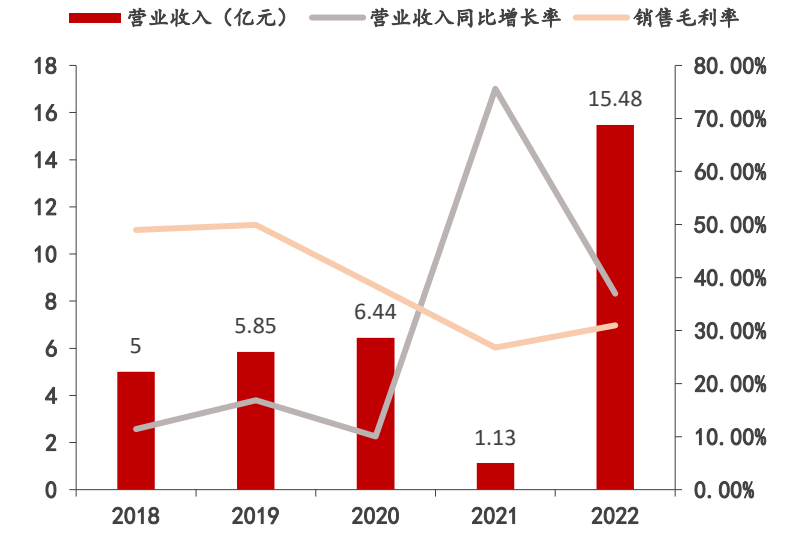


资料来源：Wind，公司年报，国投证券研究中心

6.9. 恒铭达

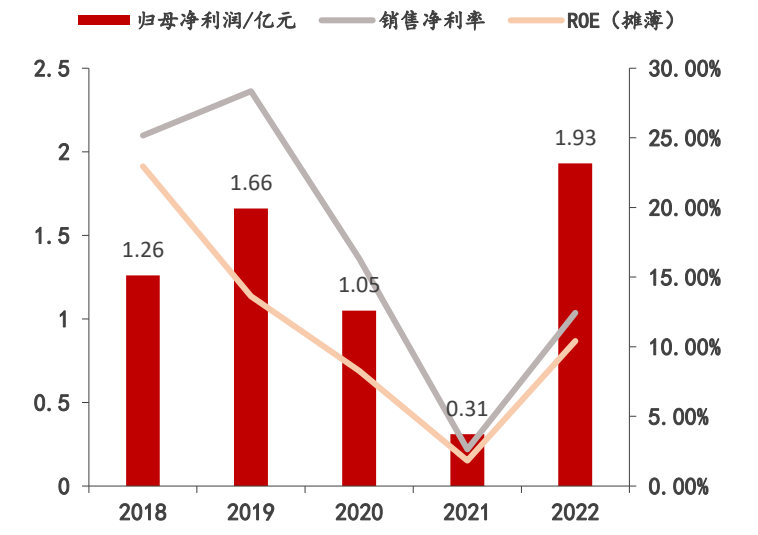
作为专业从事消费电子产品高附加值精密功能性器件的科技企业，公司具备显著的技术研发优势、丰富的生产经验、创新的生产工艺以及高精密与高品质的产品。公司的能力不仅仅局限于产品供应，而是涵盖了设计研发、材料选型、产品试制和测试、批量生产、及时配送及后续跟踪服务等一系列综合解决方案，提供了全方位的客户支持。作为国家高新技术企业，恒铭达拥有 6 项发明专利和 27 项实用新型专利，这些成就反映了其在自主研发和创新方面的扎实实力。

图38. 恒铭达营收及同比增速 (单位:亿元)



资料来源：Wind，公司年报，国投证券研究中心

图39. 恒铭达归母净利润及同比增速 (单位:亿元)



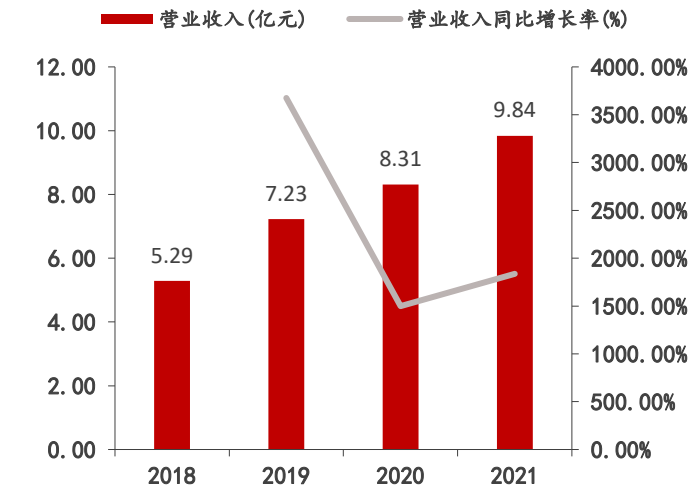
资料来源：Wind，公司年报，国投证券研究中心

6.10. 华丰科技

高速连接器，作为现代电子设备间信号传输的关键组件，其设计和功能的优势日益凸显，其具备的高速传输性、低信号损耗、低串扰性能、高密度设计，能支持高达数百 Gbps 的传输速度，对于 AI 服务器、GPU&CPU 等高带宽需求的设备来说至关重要。

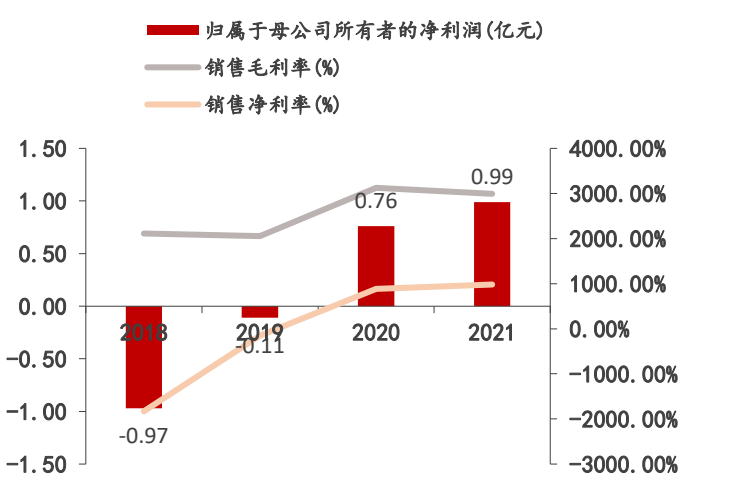
公司专注于光电连接器和线缆组件的研发、生产和销售，并向客户提供全面的系统解决方案。华丰科技以技术创新为驱动力，深耕于通讯、防务、工业等多个重要行业。其产品广泛应用于通讯、航空、航天、船舶、电子设备、核电、新能源汽车、轨道交通等关键领域。

图40. 华丰科技营收及同比增速



资料来源：wind，国投证券研究中心

图41. 华丰科技归母净利润及利润率情况



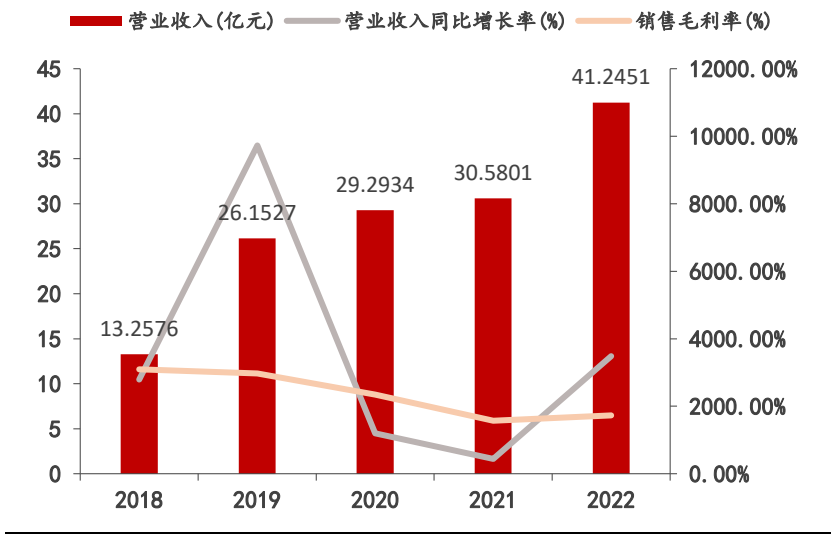
资料来源：wind，国投证券研究中心

6.11. 飞荣达

自 1993 年成立以来，公司在电子辅料产品生产领域取得了显著发展。自 2000 年开始，公司着手开发国际主流品牌的电磁屏蔽和导热材料与相关散热解决方案，逐步扩展其产品线至金属屏蔽器件、导电布衬垫、导热界面器件等，拓展应用至计算机和手机等高端领域。据公司公告与 2022 年年报，在通信基站领域，通过与中兴通讯的合作，公司成功开发并试制了 3D VC 技术样机，标志着 5G 基站首次采用这一先进的散热技术。在服务器领域，飞荣达向包括华为和超聚变在内的多个知名客户提供了一系列散热与电磁屏蔽解决方案和产品。其产品线涵盖单相和两相液冷模组、轴流风扇以及专门设计的散热器等，以满足客户多样化的散热需求。

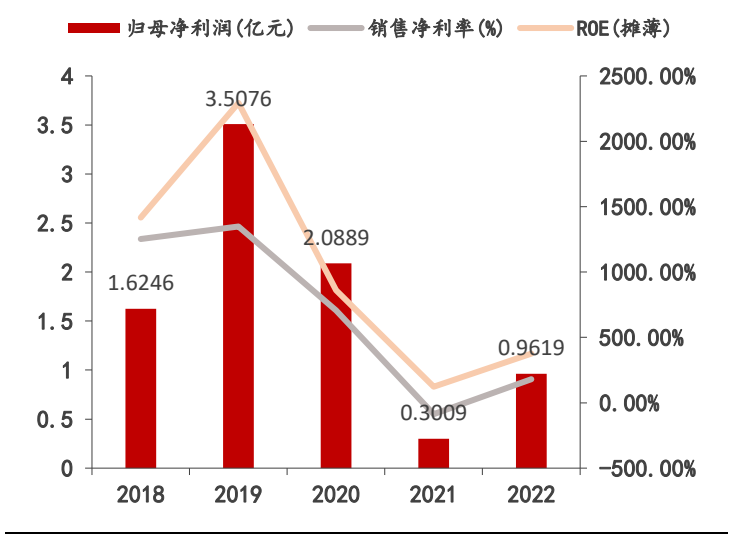
公司不仅拥有华为、中兴、微软等国内外大型企业客户，还成功扩展至 Facebook 和 Google 等全球知名企业。这一广泛的客户基础证明了其产品的实力和市场认可度。飞荣达的优势还体现在其完整的散热产业链布局上，公司不仅在散热器件和材料领域具有强劲竞争力，而且上下游一体化的产业链布局进一步增强了公司在整个散热行业的综合竞争优势。随着液冷散热市场的不断扩大及市场潜力逐步释放，飞荣达有望获得显著的市场机遇。

图42. 飞荣达营收及同比增速 (单位:亿元)



资料来源：Wind，公司年报，国投证券研究中心

图43. 飞荣达归母净利润及同比增速 (单位:亿元)

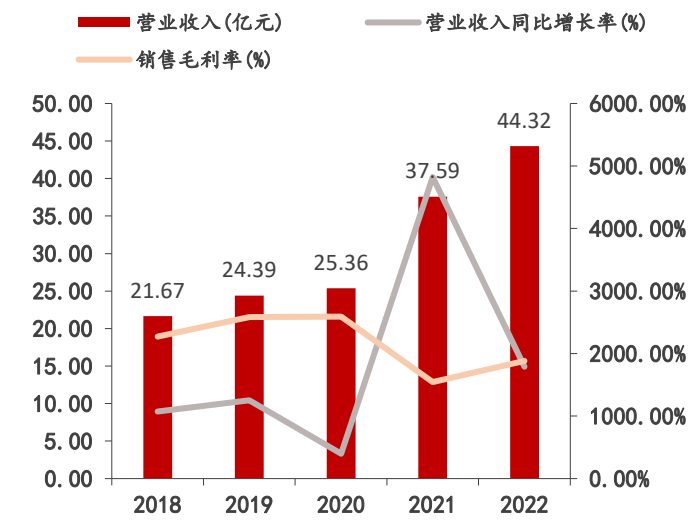


资料来源：Wind，公司年报，国投证券研究中心

6.12. 世运电路

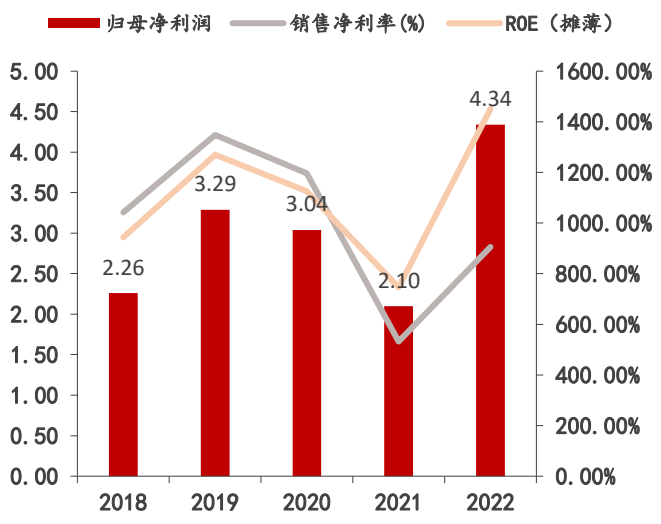
世运电路作为国内 PCB 行业的大型先进企业之一，拥有从单面板到 24 层板、金属基板、高密度互联（HDI）、软板和软硬结合板等多种线路板产品。公司目前正建设“年产 300 万平方米线路板新建项目”，分三期建设，其中一期项目已于 2022 年投产。预计该项目全部达产后，公司整体产能将增加至 700 万平方米，为公司未来在新能源汽车、光伏储能和人工智能等领域的技术市场转型提供强大的产能支持。

图44. 世运电路营收及同比增速（单位:亿元）



资料来源：Wind，公司年报，国投证券研究中心

图45. 世运电路归母净利润及同比增速（单位:亿元）

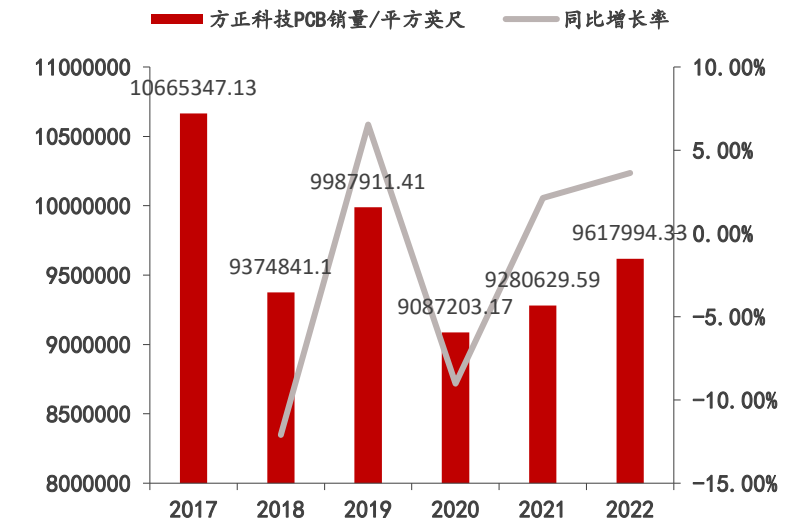


资料来源：Wind，公司年报，国投证券研究中心

6.13. 方正科技：

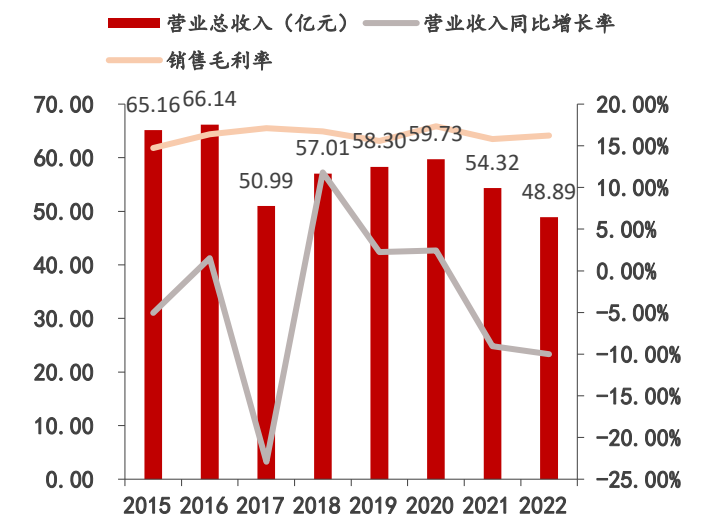
方正科技成立于 1993 年，目前已经成为国内首屈一指的集 PCB 产品研发、生产、制造和销售的龙头企业。公司业务涵盖 PCB 元器件——高速宽带接入——多垂直行业解决方案，致力于利用云计算、大数据分析等前沿技术为智慧城市建设提供从顶层设计到垂直行业软硬件解决方案。公司 PCB 产品客户主要集中在通信设备、通讯终端、IT 产品、工业医疗、汽车电子、消费电子等领域。

图46. 方正科技 PCB 销量及同比变化



资料来源：萝卜投研，国投证券研究中心

图47. 方正科技营收及同比增速



资料来源：Wind，国投证券研究中心

7. 风险提示

7.1. 宏观因素波动影响下游需求

下游需求与华为产业链开拓情况以及 AI 芯片生产有较强相关性。政治政策问题、芯片技术壁垒突破问题都可能导致上游制造商对应的下游目标市场需求无法达到预期。

7.2. 市场开拓不及预期

各公司在推动华为产业链各环节发展时，可能出现技术突破难度大等问题。若提供产品良率以及产量难以符合下游需求，可能导致公司市场渗透程度不及预期，对公司后续长期入局华为产业链带来困难，可能影响公司营收。

7.3. 行业竞争加剧

当前华为产业链前景广阔，是国内市场在半导体领域的重要推动力。因此更多的供应商带来了较大的竞争压力。半导体领域各龙头企业陆续布局华为产业链，如果市场需求不足预期，可能导致供大于求，相关产品价格持续走低，利润空间大幅收缩。

行业评级体系

收益评级：
领先大市 —— 未来 6 个月的投资收益率领先沪深 300 指数 10%及以上；
同步大市 —— 未来 6 个月的投资收益率与沪深 300 指数的变动幅度相差-10%至 10%；
落后大市 —— 未来 6 个月的投资收益率落后沪深 300 指数 10%及以上；
风险评级：
A —— 正常风险，未来 6 个月的投资收益率的波动小于等于沪深 300 指数波动；
B —— 较高风险，未来 6 个月的投资收益率的波动大于沪深 300 指数波动；

分析师声明

本报告署名分析师声明，本人具有中国证券业协会授予的证券投资咨询执业资格，勤勉尽责、诚实守信。本人对本报告的内容和观点负责，保证信息来源合法合规、研究方法专业审慎、研究观点独立公正、分析结论具有合理依据，特此声明。

本公司具备证券投资咨询业务资格的说明

国投证券股份有限公司（以下简称“本公司”）经中国证券监督管理委员会核准，取得证券投资咨询业务许可。本公司及其投资咨询人员可以为证券投资人或客户提供证券投资分析、预测或者建议等直接或间接的有偿咨询服务。发布证券研究报告，是证券投资咨询业务的一种基本形式，本公司可以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，并向本公司的客户发布。

免责声明

本报告仅供国投证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因为任何机构或个人接收到本报告而视其为本公司的当然客户。

本报告基于已公开的资料或信息撰写，但本公司不保证该等信息及资料的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映本公司于本报告发布当日的判断，本报告中的证券或投资标的价格、价值及投资带来的收入可能会波动。在不同时期，本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。本公司不保证本报告所含信息及资料保持在最新状态，本公司将随时补充、更新和修订有关信息及资料，但不保证及时公开发布。同时，本公司有权对本报告所含信息在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。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，一切须以本公司向客户发布的本报告完整版本为准，如有需要，客户可以向本公司投资顾问进一步咨询。

在法律许可的情况下，本公司及所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务，提请客户充分注意。客户不应将本报告为作出其投资决策的惟一参考因素，亦不应认为本报告可以取代客户自身的投资判断与决策。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议，无论是否已经明示或暗示，本报告不能作为道义的、责任的和法律的依据或者凭证。在任何情况下，本公司亦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。

本报告的估值结果和分析结论是基于所预定的假设，并采用适当的估值方法和模型得出的，由于假设、估值方法和模型均存在一定的局限性，估值结果和分析结论也存在局限性，请谨慎使用。

国投证券股份有限公司对本声明条款具有惟一修改权和最终解释权。

国投证券研究中心

深圳市
地　址：深圳市福田区福田街道福华一路 119 号安信金融大厦 33 楼
邮　编：518026
上海市
地　址：上海市虹口区东大名路 638 号国投大厦 3 层
邮　编：200080
北京市
地　址：北京市西城区阜成门北大街 2 号楼国投金融大厦 15 层
邮　编：100034