

研究所：

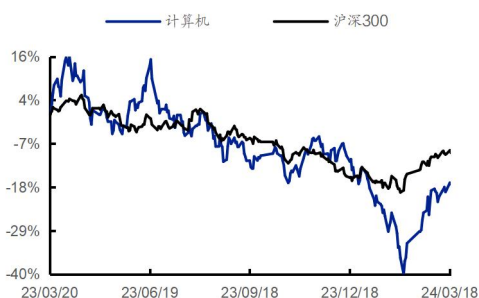
证券分析师：

刘熹 S0350523040001
liux10@ghzq.com.cn

NVIDIA B200 再创算力奇迹，液冷、光模块持续革新

——AI 算力月度跟踪（202403）

最近一年走势



行业相对表现

表现	2024/03/19		
	1M	3M	12M
计算机	17.2%	-7.2%	-18.3%
沪深 300	5.1%	7.3%	-9.6%

相关报告

《计算机行业事件点评：NVIDIA H20 或将与国产算力同步高增（推荐）*计算机*刘熹》——2024-03-03

《计算机行业月报：英伟达业绩再超预期，国产算力需求再扩大（推荐）*计算机*刘熹》——2024-02-28

《计算机事件点评：央企 AI 专题会：推动 AI 生态“国家队”加快布局（推荐）*计算机*刘熹》——2024-02-22

《AI 算力“卖水人”系列专题（1）：2024 年互联网 AI 资本开支持续提升（推荐）*计算机*刘熹》——2024-02-19

《计算机行业事件点评：Sora 震撼发布，加速 AGI 时代到来（推荐）*计算机*刘熹》——2024-02-18

投资要点：

■ AI 芯片：英伟达 B200 性能大幅提升，AI+行动提上两会议程

1) 全球：①英伟达：B200 于 GTC 2024 推出，售价约 3-4 万美元，集成 2080 亿个晶体管，采用台积电 N4P 制程，为双芯片架构，192GB HBM3E，AI 算力达 20petaFLOPS (FP4)，是 Hopper 的 5 倍；H200 预期 Q2 上市；②AMD：MI 系列加速迭代，完成产品 AI 效能组合；③Intel：Gaudi 3 预期于 2024Q3 上市，Falcon Shores 预期 2025 年发布。

2) 国产：“人工智能+ (AI+)”在 2024 年两会中首次被写进政府工作报告。英伟达首次将华为认定为人工智能芯片等多个领域的主要竞争对手。景嘉微推出景宏系列高性能智算模块及整机产品。

■ 算力设备：2024 年 AI 服务器或增三位数，英伟达 GB200 NVL72 采用液冷

1) AI 服务器：①全球：2 月，广达营收 842.93 亿新台币，同比+1.13%，2024 年下半年 AI 伺服器出货可望明显好转；纬创营收 810.97 亿新台币，环比+20.5%，同比+30.66%，AI 伺服器预估全年同比增长三位数；英业达营收 366.51 亿新台币，环比-16.95%，同比+2.27%；鸿海营收 3,525 亿元新台币，环比-32.5%、同比-12.3%，2024 年 AI 服务器市场份额目标是 40%。②内地：华为官网更新硬件合作伙伴级别名单，其中中华鲲鹏振宇+超聚变升级为华为战略级伙伴（最高级）。华为公司董事 ICT 产品与解决方案总裁杨超斌表示，鲲鹏、昇腾逐步为数智化转型升级首选算力。

2) 散热：①全球：英伟达 GB200 NVL72 使用液冷机柜。预计 2024 年液冷服务器的占比将迅速提升，3D VC 散热方案也有望持续增长；②内地：飞荣达、曙光数创均表示散热产品量产交付、高速增长；英维克用于 5MWh 储能系统及工商业储能融合液冷机组将亮相。

3) 交换机：NVIDIA 发布为大规模 AI 量身定制的全新网络交换 X800 系列。NVIDIA Quantum-X800 InfiniBand 网络和 NVIDIA Spectrum™-X800 以太网网络是全球首批高达 800Gb/s 端到端吞吐量的网络平台。

4) 光模块：新易盛表示加速硅光、相干光模块、1.6T 光模块等行业前沿领域研究及商用；天孚通讯表示持续信息系统升级和自动化升级、江西生产基地降本增效，产能利用率明显提升。

5) 主板：深南电路表示，无锡基板二期工厂、广州封装基板项目一

期工厂已先后连线,目前均处于产能爬坡阶段。华擎表示,预期今年 AI 伺服器将加入贡献,将推升伺服器业务较去年明显增长。

■ **GPU 产业链:台积电 3nm/4nm 产能满载,Meta 新建集群含超 4.9 万块 H100**

1) COWOS: 英伟达将采用台积电 3/4nm 制程,强劲订单推动先进制程产能满载。据财联社,台积电 3 月追加 CoWoS 设备订单,交机时间预计今年四季度,预计 2024 年底台积电 CoWoS 月产能或超过 4 万片。

2) 互联网需求旺盛: Meta 公布了两座新的数据中心集群,内含超 4.9 万块英伟达 H100 GPU,专门训练 Llama3。AMD 为 Instinct MI300 系列加速器谋发展,将以较低价格供货给微软,促进双方新一轮长期合作。

行业评级及投资策略: 大模型带动 AI 算力需求增长,算力产业链中的 AI 芯片厂商、上游封装厂商与下游服务器整机厂商有望持续受益。维持对计算机行业“推荐”评级。

相关公司: **1) 服务器整机:** 工业富联、浪潮信息、中科曙光、华勤技术、中国长城、高新发展、神州数码、烽火通信、拓维信息、纬创、广达、英业达、纬颖、超微电脑。**2) 服务器组件:** ①**AI 芯片:** 海光信息、寒武纪、龙芯中科、景嘉微、英伟达、AMD、Intel; ②**散热:** 飞荣达、曙光数创、英维克、同方股份、申菱环境、高澜股份、奇鋆科技、双鸿、VERTIV; ③**主板:** 沪电股份、深南电路、胜宏科技、技嘉、华擎。**3) 光模块:** 天孚通信、中际旭创、新易盛、光迅科技、华工科技。**4) 数据中心:** 奥飞数据、光环新网、宝信软件、数据港、电科数字。

风险提示: 宏观经济影响下游需求,信创政策不及预期,市场竞争加剧,中美博弈加剧,相关公司业绩不及预期等,各公司并不具备完全可比性,对标的相关资料和数据仅供参考。

内容目录

1、 AI 芯片：英伟达 B200 性能优越，“AI+”行动提上两会议程	5
1.1、 全球 AI 芯片：英伟达 B200 FP4 性能约为 Hopper 的 5 倍，AMD/Intel 加速产品升级	5
1.2、 国产 AI 芯片：“AI+”行动提上两会议程，英伟达将华为定义为公司竞争对手	8
2、 服务器及产业链：广达/纬创预期 AI 服务器 2024 年高速增长，液冷产品加速迭代升级	10
2.1、 AI 服务器：纬创预期 2024 年 AI 服务器业务三位数增长，华为合作伙伴中标多项服务器集采项目	10
2.2、 服务器散热：英伟达 GB200 NVL72 采用液冷系统，液冷有望高增长	13
2.3、 交换机：英伟达发布 X800 系列适配 AI 计算，联想/超微等预期 2025 年供货	15
2.4、 光模块：1.6T 研发加速，市场空间广阔	16
2.5、 主板：厂商产能持续爬坡，2024 年预期增长	16
3、 GPU 产业链：台积电 3nm/4nm 产能满载，Meta 新建集群含超 4.9 万块 H100	17
3.1、 COWOS：台积电 3nm/4nm 产能满载	17
3.2、 互联网需求旺盛：Meta 新建集群含超 4.9 万块 H100	18
4、 相关公司	19
5、 风险提示	20

图表目录

图 1: 未来 AMD 芯片发展路径	7
图 2: Intel AI 芯片发展路径	8
图 3: Gaudi 3 相对 Gaudi2 升级情况	8
图 4: 台股各服务器厂商月度营收及同比情况	11
图 5: 昇腾生态伙伴建设 (2024.03)	11
图 6: 台股各散热厂商月度营收及同比情况	13
图 7: 2023~2024 年全球 CSP 高阶 AI 服务器需求	19
图 8: Meta 新建两座数据中心集群	19
图 9: 服务器产业链相关公司	20
表 1: Blackwell 平台的核心技术及主要产品	5
表 2: 英伟达 HGX B200 与 HGX B100 的系统参数	6
表 3: 英伟达发展应用端, 包括机器人、自动驾驶、推理微服务等	6
表 4: 2024 “两会” 关于国产 AI 芯片提案	9
表 5: 多家国产芯片厂商完成适配与商业化应用	9
表 6: 广达/纬创预期 AI 服务器 2024 年增长快速	10
表 7: 在华为合作伙伴大会上多家华为合作伙伴推出新品	12
表 8: 散热大厂加速布局液冷散热, 东南亚基地产能扩增	14
表 9: 液冷产品升级迭代加速, 智算中心建设迎来新趋势	14
表 10: 英伟达发布 X800 系列交换机, 并提供全套软件方案	15
表 11: 光模块厂商持续扩产, 1.6T 产品研发加速	16
表 12: 基板厂商产能持续爬坡, 2024 年预期恢复增长	17
表 13: 台积电 3nm、4nm 产能近满载, 预计 Q4 CoWos 月产能超 4 万片	18

1、AI 芯片：英伟达 B200 性能优越，“AI+”行动上两会议程

1.1、全球 AI 芯片：英伟达 B200 FP4 性能约为 Hopper 的 5 倍，AMD/Intel 加速产品升级

(1) 英伟达：B200 于 GTC2024 推出，FP4 性能约为 Hopper 的 5 倍

Blackwell GPU 计算性能优越。3 月 18 日，NVIDIA GTC 2024 召开，英伟达 CEO 黄仁勋发布了 B200 GPU：集成 2080 亿个晶体管，采用台积电 N4P 制程，为双芯片架构，两块小芯片间互联速度高达 10TBps；配备 192GB HBM3E，存储带宽 8TB/s，AI 算力达到 20petaFLOPS(FP4 精度)，而 Hopper 是 4 petaFLOPS。

B200 预计 2024 年晚些推出、售价 3-4 万美元，H200 预期 Q2 上市。据财联社，B200 GPU 售价或为 3-4 万美元，英伟达称，包括亚马逊云科技、戴尔科技、谷歌、Meta、微软、OpenAI、甲骨文、特斯拉和 xAI 将计划采用 Blackwell 产品。

表 1：Blackwell 平台的核心技术及主要产品

GPU	具体内容
Blackwell GPU	B200: 集成 2080 亿个晶体管，采用台积电 N4P 制程，两块小芯片的互联速度高达 10TBps。配备 192GB HBM3E，存储带宽 8TB/s，AI 算力达到 20petaFLOPS(FP4 精度)，而 Hopper 是 4 petaFLOPS GB200: 两个 B200 GPU 与 Grace CPU 结合，在标准的 1750 亿参数 GPT-3 基准测试中，GB200 的性能是 H100 的 7 倍，提供的训练算力是 H100 的 4 倍
第二代 Transformer 引擎	将新的微张量缩放支持和先进的动态范围管理算法与 TensorRT-LLM 和 NeMo Megatron 框架结合，使 Blackwell 具备在 FP4 精度的 AI 推理能力，可支持 2 倍的计算和模型规模，能在将性能和效率翻倍的同时保持混合专家模型的高精度。
第五代 NVLink 互连	第五代 NVLink 互连则是将多个 Blackwell GPU 组合起来的重要工具。它与传统的 PCIe 交换机不同，NVLink 带宽有限，可以在服务器内的 GPU 之间实现高速直接互连。目前第五代 NVLink 可每个 GPU 提供了 1.8TB/s 双向吞吐量，确保多达 576 个 GPU 之间的无缝高速通信。
RAS 可靠性引擎	基于 AI 实现，Blackwell 透过专用的可靠性、可用性和可维护性 (RAS) 引擎，可增加智慧复原能力，及早辨认出可能发生的潜在故障，尽可能缩短停机时间。
Secure AI 安全 AI 功能	负责提供机密运算功能，同时 Blackwell 也是业界第一款支持 EE-I/O 的 GPU，它可以在不影响性能的前提下，维护你的数据安全，这对于金融、医疗以及 AI 方面有极大作用。
专用解压缩引擎	资料分析和资料库工作流程此前更多是仰赖 CPU 进行运算。如果放到 GPU 中进行则可大幅提升端对端分析的效能，加速创造价值，同时降低成本。
系统	具体内容
DGX B200	配备 8 个 NVIDIA Blackwell GPU 与第五代 NVLink 互连，训练性能是前几代产品的 3 倍，推理性能是前几代产品的 15 倍；DGX B200 可以处理各种工作负载，包括大型语言模型、推荐系统和聊天机器人。
GB200 NVL72	集成 36 个 Grace CPU 和 72 个 Blackwell GPU，采用液冷机柜，拥有 30TB 高速内存，FP8/FP6 Tensor Core 性能 720PFLOPS(训练)，或 FP4 Tensor Core 性能 1,440 PFLOPS(推理)。与相同数量的 H100 Tensor Core GPU 相比，GB200 NVL72 在 LLM 推理工作负载性能最多可提升 30 倍，成本和能耗最多可降低 25 倍。
DGX GB200 Superpod	全新 DGX SuperPOD 是一台完整的数据中心级 AI 超级计算机。DGX SuperPOD 采用新型高效液冷机架级扩展架构，由 8 个或以上的 DGX GB200 系统构建而成，通过 NVIDIA Quantum InfiniBand 网络连接，可扩展到数百万个 GB200 超级芯片。在 FP4 精度下可提供 11.5 exaflops 的 AI 超级计算性能和 240 TB 的快速显存，且可通过增加机架来扩展性能。

资料来源：英伟达官网，PConline 太平洋科技，芯智讯，OSC 开源社区，国海证券研究所

表 2: 英伟达 HGX B200 与 HGX B100 的系统参数

	HGX B200	HGX B100
Blackwell GPUs	8	8
FP4 Tensor Core	144 PetaFLOPS	112 PetaFLOPS
FP8/FP6/INT8	72 PetaFLOPS	56 PetaFLOPS
Fast Memory	Up to 1.5 TB	Up to 1.5TB
Aggregate Memory Bandwidth	Up to 64 TB/s	Up to 64 TB/s
Aggregate NVLink Bandwidth	14.4 TB/s	14.4 TB/s
单 GPU 性能		
FP4 Tensor Core	18 petaFLOPS	14 petaFLOPS
FP8/FP6 TensorCore	9 petaFLOPS	7 petaFLOPS
INT8 Tensor Core	9 petaOPS	7 petaOPs
FP16/BF16 TensorCore	4.5 petaFLOPS	3.5 petaFLOPS
TF32 Tensor Core	2.2 petaFLOPS	1.8 petaFLOPs
FP64 Tensor Core	40 teraFLOPS	30 teraFLOPS
GPU memory Bandwidth	Up to 192 GB HBM3e Up to 8 TB/s	
Max thermal design power (TDP)	1000W	700W
Interconnect	NVLink: 1.8TB/s PCIe Gen6: 256GB/s	NVLink: 1.8TB/s PCIe Gen6: 256GB/s
Server options	NVIDIA HGX B200 partner and NVIDIA-Certified Systems with 8 GPUs	NVIDIA HGX B200 partner and NVIDIA-Certified Systems with 8 GPUs

资料来源：英伟达，半导体行业观察，国海证券研究所

英伟达发力应用端。1) 机器人：推出了 GR00T 项目，为人形机器人提供了通用基础模型，也对 Isaac 机器人平台进行重大更新；**2) 自动驾驶：**集中式车载计算平台 DRIVE Thor 将搭载专为 Transformer、大语言模型（LLM）和生成式 AI 工作负载而打造的全新 Blackwell 架构；**3) 推理微服务 NIM：**英伟达提供预训练好的 AI 模型并开放 API（应用程序接口），再由行业客户开发应用，以简化企业自己开发生成式 AI 应用的成本，每个 GPU 每年收取费用 4500 美元。

表 3: 英伟达发展应用端，包括机器人、自动驾驶、推理微服务等

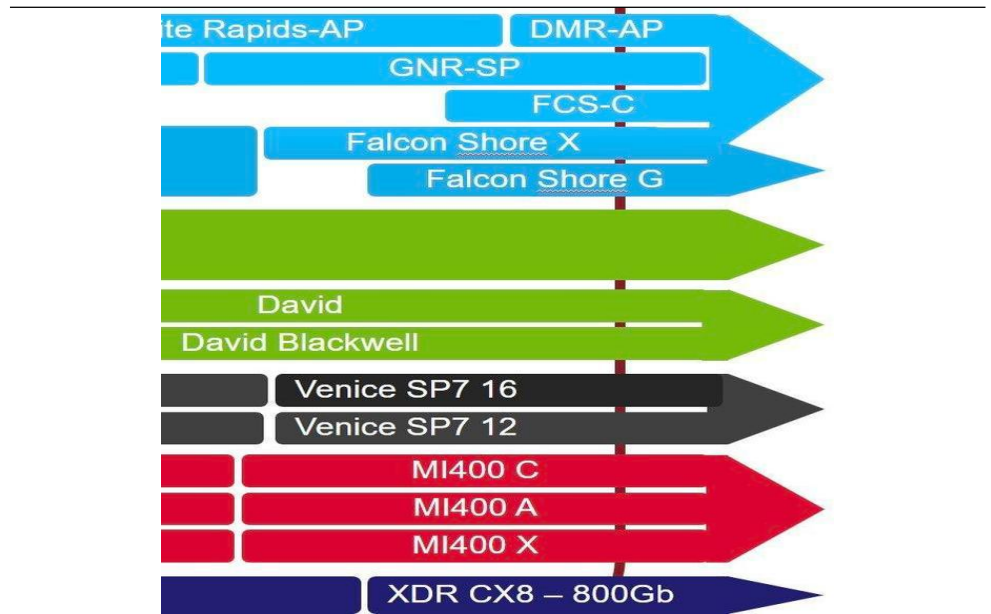
提案人	
机器人	<ul style="list-style-type: none"> ● GR00T 项目：为人形机器人提供了通用基础模型，旨在推动机器人和具象化人工智能领域的突破。GR00T 将作为机器人的智能之躯，使其能够学习各种技能以解决各种任务 ● Isaac 机器人平台：重大更新，包括引入了新的机器人训练模拟器、Jetson Thor 机器人计算机、生成式 AI 基础模型以及 CUDA 加速的感知和操作库。 ● Jetson Thor：基于 Thor 芯片的新型人形机器人计算机，提供每秒 800 万亿次 8 位浮点运算 AI 性能，可以运行 GR00T 等多模态生成式 AI 模型，并大大简化设计和集成工作
自动驾驶	<ul style="list-style-type: none"> ● 集中式车载计算平台 DRIVE Thor 将搭载专为 Transformer、大语言模型（LLM）和生成式 AI 工作负载而打造的全新 Blackwell 架构。 ● 多家头部电动汽车制造商在 GTC 上展示了其搭载 DRIVE Thor 的下一代 AI 车型，既包括比亚迪、广汽埃安、小鹏、理想汽车和极氪等众多中国车企，也包括了文远知行等自动驾驶平台公司。
推理微服务 NIM	<ul style="list-style-type: none"> ● NIM 可以更容易地使用旧的英伟达 GPU 进行推理，并允许公司继续使用他们已经拥有的数亿个英伟达 GPU。该公司的策略是让购买英伟达服务器的客户注册英伟达企业版，每个 GPU 每年收取费用 4500 美元。 ● 量子计算领域：英伟达宣布推出云量子计算机模拟微服务，帮助研究人员和开发人员在化学、生物学、材料科学等科学领域的量子计算研究，该服务基于开源 CUDA-Q 量子计算平台。 ● 医药领域：英伟达宣布旗下包括 Parabricks、MONAI、NeMo™、Riva、Metropolis，现已通 CUDA-X 微服务提供访问，以加速药物研发、医学影像、基因组学分析等医疗工作流程。

资料来源：界面新闻微信公众号，国海证券研究所

(2) AMD: MI 系列产品加速迭代, 完成全产品系列 AI 效能组合

MI300 改版/ MI400 即将推出。据 IT 之家 2 月 23 日消息, AMD 将推出换用 HBM3e 的 AI 加速器 MI300 改版, 以低价与竞品英伟达 B100 竞争, 其内存在速度上进一步提升; 而下一代 Instinct MI400 加速器将于 2025 年发布, 包含 Mercury / Venus / Earth 三个代号、X / A / C 三种变体类别, 可实现从云端到边缘、PC 和智能终端设备的新体验和突破性 AI 功能。

图 1: 未来 AMD 芯片发展路径



资料来源: IT 之家

MI300 中国特供版芯片 MI309 或推出, 但仍需关注出口禁令限制情况。据 3 月 5 日钛媒体, AMD 即将向中国客户销售 MI300 系列“中国特供版”AI 芯片, 被称为 MI309, 但是, 美国政府目前不同意这款特供芯片出口到中国。截至 3 月 5 日, AMD、商务部拒绝发表评论, 尚不清楚 AMD 是否依然在申请许可。

(3) Intel: Gaudi 3 预期于 2024Q3 上市, Falcon Shores 预期 2025 年发布

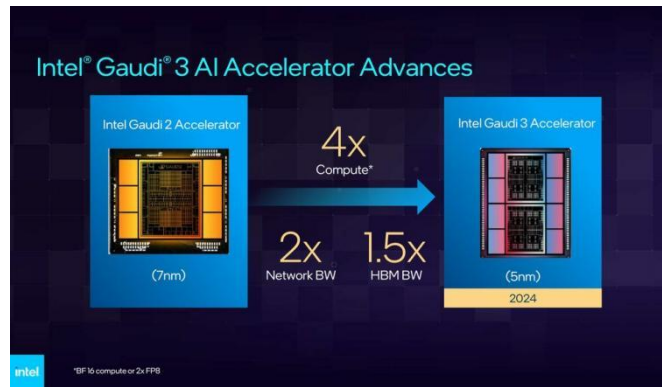
英特尔将 **Gaudi 3** 和 **Falcon Shores** 定位为后续产品, 两者都属于同一条产品线。据超能网, Gaudi 3 将改用 5nm 工艺制造, 已进入实验室进行验证, Gaudi 3 计划在 2024 年第三季度全面上市。硬件方面, Gaudi 3 加速器将采用与 Gaudi 2 相同的高性能架构, 不过计算能力是其 4 倍, 网络带宽是其 2 倍, HBM 内存带宽是其 1.5 倍。Falcon Shores 是数据中心 GPU, 采用多芯片模块化设计, 同时会加入“可扩展的 I/O 设计”, Falcon Shores 计划 2025 年发布。

图 2: Intel AI 芯片发展路径



资料来源: 超能网

图 3: Gaudi 3 相对 Gaudi2 升级情况



资料来源: 超能网

英特尔设立独立 **FPGA 业务部门**, 新工厂极大提振 **AI 芯片产能**。据 3 月 1 日 IT 之家, 英特尔公司宣布以独立发展模式, 正式成立 **FPGA90** (现场可编程门阵列) 公司 **Altera**, 并推出了包括 **Agilex 9**、**Agilex 7 F** 系列和 **I** 系列、**Agilex 5** 和 **Agilex 3** 等产品。其中, **Agilex 5** 现已广泛上市, 是唯一注入人工智能的 **FPGA** 结构, 每瓦性能是同类产品的 **1.6 倍**, 面向嵌入式和边缘应用。

1.2、国产 AI 芯片：“AI+”行动提上两会议程，英伟达将华为定义为公司竞争对手

人工智能+首次写进政府工作报告。在十四届全国人大二次会议上, 李强总理在 2024 年政府工作报告中提出, 深化大数据、人工智能等研发应用, 开展“人工智能+”行动, 从以往的“互联网+”思维模式, 转变至未来的“人工智能+”思维模式。“人工智能+”在 2024 年两会中首次被写进政府工作报告, 意味着国家将加强顶层设计, 加快形成以 **AI** 为引擎的新质生产力。

表 4: 2024 “两会” 关于国产 AI 芯片提案

提案人	提案 / 建言	要点
邓中翰 全国政协委员 中国工程院院士	AI 时代精准支持芯片产业高质量发展	AI 时代竞争靠的是算力，关键在于芯片。建议我国相关部门围绕智能计算、智能感知等新型领域，加快开展自主标准制定和关键技术核心芯片的攻关研发，以自主标准优势，结合发挥政策优势，推动形成产业优势。
湛志华 麒麟软件董事长	关于夯实人工智能发展算力底座，加速推动国产操作系统发展	通过基础软件先行的方式，深入挖掘国产 AI 芯片算力潜力，依托专业操作系统龙头企业，与各 AI 芯片厂商深度合作，研制能够兼容各类国产 AI 芯片和训练/推理框架的智算操作系统，实现智算集群算力异构，灵活调度智算集群算力，完成各种训练和推理任务。
曹鹏 京东技术委员会主席	《加速发展新质生产力，推动国产自研技术和产业发展深度融合》	通过政策鼓励国产化 GPU 适配国产的算力调度软件，建设自主可控的智算基础，支撑行业智能化发展。
李家杰 香港恒基兆业集团主席、香港中华煤气有限公司主席	《关于加快突破芯片产业发展瓶颈的提案》	加大对国产芯片优质企业的支持力度，建议推动 RISC-V 开源指令集成为行业标准，并鼓励国内相关企业参与到标准规范的讨论与制定。
苗伟 中兴通讯高级副总裁	加大算力基础设施建设，培育新业态新模式	在算力基础设施建设方面，苗伟建议设立跨区域协同的算力建设扶持政策，引导社会资本和金融机构积极参与算力基建和技术研发，同时加强对国内 GPU 厂商的支持和统筹引导等，以推动算力基础设施的高质量发展，提升智能算力供给能力，进一步促进数字经济蓬勃发展。

资料来源：中国经营网，antpedia，新浪财经，自主可控新鲜事，国海证券研究所

国产 GPU：英伟达将华为定义为公司竞争对手，景嘉微推出景宏系列 AI 芯片。据 3 月 14 日 IT 之家，海思 5nm 麒麟处理器在东莞、北京等地的验证情况良好，预计 8 月开始大规模交付；根据快科技，在英伟达近日向美国证券交易委员提交的正式文件中，首次将华为认定为人工智能芯片等多个领域的主要竞争对手。3 月 14 日，景嘉微推出景宏系列，是面向 AI 训练、AI 推理、科学计算等应用领域的高性能智算模块及整机产品。

表 5: 多家国产芯片厂商完成适配与商业化应用

公司	具体内容
华为海思	● 据 3 月 14 日 IT 之家，海思 5nm 麒麟处理器在东莞、北京等地的验证情况良好，预计 8 月开始大规模交付。据路透社，英伟达将华为定位为一家设计自己的硬件和软件以改进人工智能计算的云服务公司，并表示，华为在供应图形处理单元（GPU）、中央处理单元（CPU）和网络芯片等人工智能芯片方面与公司存在竞争。
海光 DCU	● 2023 年公司营收 601,199.90 万元，同比+17.30%；实现归母净利润 126,244.13 万元，同比+57.11%。 ● 2023 年公司围绕通用计算市场，持续保持高强度研发投入，通过技术创新，进一步提升了产品性能，得到客户充分认可，在毛利率方面有所提升，实现了业绩的持续增长。
寒武纪	● 3 月 4 日公司表示，2022 年思元 370 系列产品在多家头部企业完成产品导入。2023 年上半年，公司云端产品在互联网、运营商、金融行业及客户中进行了广泛的业务部署与落地。基于云端产品的优势，针对最近兴起的大模型领域，优化了公司产品在 AIGC 及大语言模型领域的性能，并与多个行业客户及 ISV 推动了技术和产品合作。 ● 3 月 1 日公司表示，思元 370 智能芯片及加速卡是基于公司第四代智能处理器微架构的。
景嘉微	● 3 月 14 日，公司表示景宏系列是公司推出的面向 AI 训练、AI 推理、科学计算等应用领域的高性能智算模块及整机产品，支持 INT8、FP16、FP32、FP64 等混合精度运算，支持全新的多卡互联技术进行算力扩展，适配国内外主流 CPU、操作系统及服务器厂商，能够支持当前主流的计算生态、深度学习框架和算法模型库，大幅缩短用户适配验证周期。

资料来源：iFinD，公司公告，快科技，路透社，IT 之家，国海证券研究所

2、服务器及产业链：广达/纬创预期 AI 服务器 2024 年高速增长，液冷产品加速迭代升级

2.1、AI 服务器：纬创预期 2024 年 AI 服务器业务三位数增长，华为合作伙伴中标多项服务器集采项目

(1) 全球服务器：广达/纬创预期 AI 服务器 2024 年增长快速

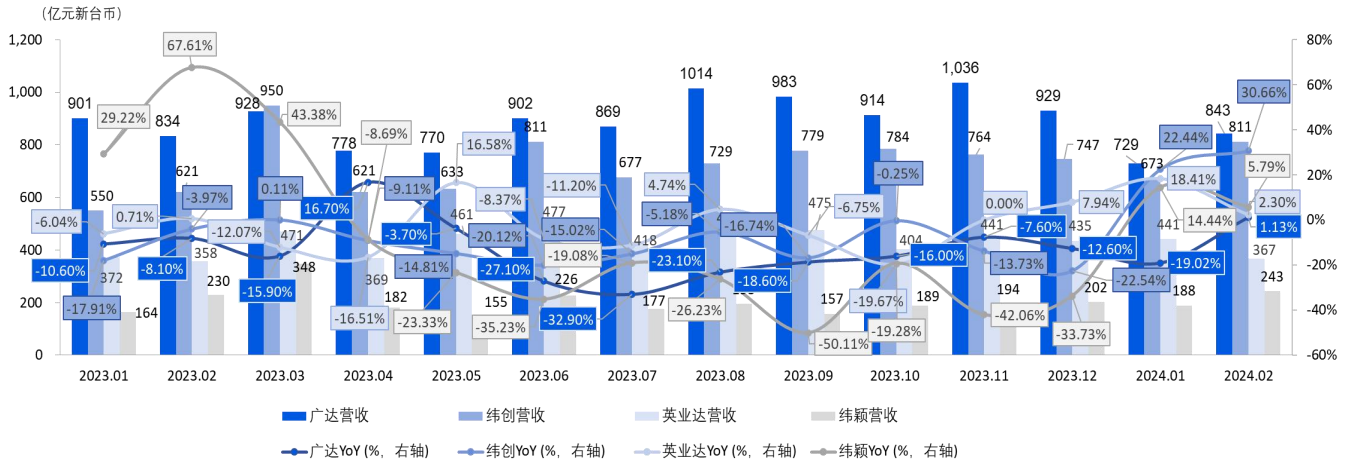
2024 年 2 月，台股服务器主要厂商 AI 服务器预期仍然乐观。广达表示 2024 年下半年 AI 伺服器出货可望明显好转，2024 年服务器业务增长将推动云端事业营收成为最大产品线。纬创表示 AI 伺服器业务展望乐观，预估全年同比增长达三位数。具体看：

表 6：广达/纬创预期 AI 服务器 2024 年增长快速

公司	详细内容
广达	<ul style="list-style-type: none"> ● 2024 年 2 月，营收达 842.93 亿新台币，同比+1.13%；前两月营收 1,572.3 亿新台币，同比-9.3%。展望 2024，随着台积电先进封装产能陆续开出，上游供给缺口可望逐步纾解，广达 2024 年下半年 AI 伺服器出货可望明显好转，推动云端事业营收占比冲破四成，有望超越笔电，成为广达最大产品线。 ● 根据 2 月 29 日财联社，英伟达 H200 芯片将于第二季开始出货，四大云端服务供应商积极争抢。Google 主力代工厂广达首批搭载 H200 的 AI 服务器已开案，预计 2024Q3 完成后段测试并量产。供应链透露，广达此次量产 H200 的 AI 服务器，订单量至少数百机柜起跳，出货单价是一般传统服务器机柜 3-4 倍。
纬创	<ul style="list-style-type: none"> ● 2024 年 2 月，营收 810.97 亿新台币，环比+20.5%，同比+30.66%，创历年同期新高；前两月累计营收 1,483.97 亿新台币，同比+26.8%；AI 伺服器需求火热，出货量将持续升温。展望 2024，因 3 月季底拉货效应，纬创推估 3 大产品线出货将环比增长达两位数，AI 伺服器业务展望乐观，预估全年同比增长达三位数。
英业达	<ul style="list-style-type: none"> ● 2024 年 2 月，营收 366.51 亿新台币，环比-16.95%，同比+2.27%。累计前两月营收 807.79 亿元，同比+10.53%。据艾邦加工展，英业达服务器因出货递延，2024Q1 营收或将环比持平，若图形处理器（GPU）供给较佳，伺服器营收有望环比上升。
纬颖	<ul style="list-style-type: none"> ● 2024 年 2 月，营收 243.26 亿元新台币，环比+29.48%，同比+5.79%。
鸿海	<ul style="list-style-type: none"> ● 2024 年 2 月，营收 3,525 亿元新台币，环比-32.5%、同比-12.3%。累计前两月营收 8,746 亿新台币，同比-17.67%。 ● 分业务来看：①云及网络产品：受益于强劲的客户拉动，2 月份营收同比大幅增长；②零组件及其他产品：智能消费电子相关零组件出货量增加，但因非核心业务减少，营收大致持平。③计算产品：由于 PC 市场需求不温不火，2 月份营收同比小幅下滑。④智能消费电子：由于 2 月工作日减少，营收同比下降。 ● 展望 2024 年，鸿海表示预计第一季度营收同比下降，预计 2024 财年营收同比增长，并预计 2024 年芯片业务销售额将超过 1000 亿元新台币；2024 年 AI 服务器市场份额目标是 40%。

资料来源：各公司官网，联合新闻网，雅虎，chinatimes，艾邦加工展，优分析，财联社，和讯网，国海证券研究所

图 4：台股各服务器厂商月度营收及同比情况



资料来源：各公司官网，国海证券研究所

(2) 内地服务器：华鲲振宇+超聚变为华为昇腾战略级伙伴

华鲲振宇和超聚变升级为华为战略级伙伴。3月2日，华为官网更新了硬件合作伙伴级别名单，其中华鲲振宇+超聚变升级为华为战略级伙伴（最高级），有且仅有两家。超聚变更是连跳三档，从认证级直接跳到战略级，创造了华为合作商的历史。神州数码为领先级合作伙伴，此外认证级、优先级伙伴共 10 家。

图 5：昇腾生态伙伴建设（2024.03）



资料来源：华为昇腾官网

华为预计昇思将成为 2024 年中国开源 AI 框架的首选。3 月 15 日，华为公司董事 ICT 产品与解决方案总裁杨超斌表示，鲲鹏、昇腾逐步成为数智化转型升级的首选算力；昇思成为成长最快的开源 AI 框架，2023 年市场份额增速第一，预计 2024 年成为中国首选。2023 年昇腾的模型和算子覆盖率、鲲鹏的应用覆盖率均快速提升，2024 年将发展超过 50 家鲲鹏和昇腾的伙伴，预计 2024 年中国区 AI 训推一体机的市场空间为 168 亿元。

表 7: 在华为合作伙伴大会上多家华为合作伙伴推出新品

公司	详细内容
超聚变	<ul style="list-style-type: none"> ● 成为华为战略级合作伙伴。3 月 2 日，华为官网更新了新的硬件合作伙伴级别名单，其中超聚变连跳三档，升级为仅有的两家之一的华为战略级伙伴（最高级），创造了华为合作商的历史。 ● 中标招商银行信创 AI 服务器大单。3 月 1 日，招商银行股份有限公司总行发布《信创 AI 服务器项目》采购结果公告。其中，超聚变全资子公司昆仑技术成功中选。
华鲲振宇 (高新发展)	<ul style="list-style-type: none"> ● 华鲲振宇推出天智系列大模型推理服务服务器：天智推理服务器 HuaKun AT3500 G3、天智推理服务器 HuaKun AT9508 G3、天智推理服务器 HuaKun AT3510 G4。其中，HuaKun AT9508 G3 是鲲鹏+昇腾生态唯一支持 10 卡双宽或 20 卡单宽的服务器，剑指业内最高规格，受到宇信科技等伙伴和客户的高度关注。
神州鲲泰 (神州数码)	<ul style="list-style-type: none"> ● 2024 年神州鲲泰获评“鲲鹏展翅领先级”“昇腾万里领先级”整机硬件伙伴。神州数码携多款产品及解决方案亮相盛会，重磅发布神州鲲泰问学一体机。 ● 中标厦门税务“新电子税务局”建设项目。该项目主要采用神州鲲泰 KunTai R722 服务器，实现全“鲲鹏”算力底座交付。
湘江鲲鹏 (拓维信息)	<ul style="list-style-type: none"> ● 重磅发布兆瀚 RA5300-B AI 推理服务器新品，充分展示全栈国产技术实力和应用成果。 ● 拓维信息系统股份有限公司副总裁柏丙军表示，过去三年，与华为合作的业务规模快速增长，年复合增长率达到 30% 以上。
长江计算 (烽火通信)	<ul style="list-style-type: none"> ● 面向通用算力场景，长江计算基于鲲鹏 920 研制的服务器产品，能匹配市场多种使用场景，且具有较强的优势和竞争力；面向智能算力场景，长江计算创新研发了算法一体机、可移动算力车、训推一体机等，可满足不同行业、不同场景的需求；面向高性能算力场景，长江计算推出的 AccelerPoD 5000 系列全液冷整机柜，能支持智能化运维，实现无人数据中心。

资料来源：各公司官方公众号，Market Monitor，iFinD，东方财富网，云头条，神州鲲泰，拓维信息，财联社，国海证券研究所

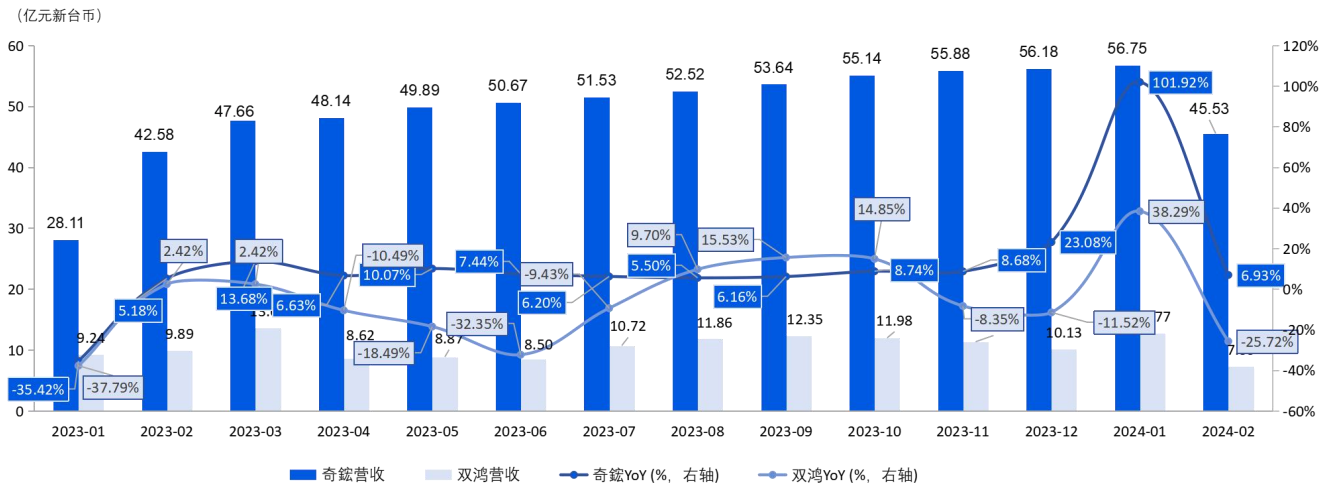
英伟达 L20/H20 方面：华勤技术看好 2024 年公司服务器业务发展。华勤技术表示，公司多产品组合在客户端可以获得全栈式解决方案的加分，特别是在人工智能服务器的核心技术上得到客户的高度认可。目前，国内互联网厂商 AI 需求在训练测和推理测均有持续增长趋势，公司基于 2023 年与国内互联网厂商较好的合作基础，2024 年有望获得更多的服务器订单量，特别是新一代 GPU 芯片平台的人工智能服务器；同时，公司快速进行行业主流的新平台芯片产品的适配，在完成相关测试认证后，根据客户的需求近期开始批量陆续出货。

2.2、服务器散热：英伟达 GB200 NVL72 采用液冷系统，液冷有望高增长

(1) 全球散热：英伟达 GB200 NVL72 采用液冷系统，散热厂商预期 2024 全年营收预期趋好

GB200 NVL72 为支持多节点与液冷的机架系统。3 月 19 日，英伟达召开 GTC2024 大会，发布了超级芯片 GB200，以及以 GB200 为核心的 GB200 NVL72 数据中心机架系统。GB200 NVL72 为支持多节点与液冷的机架系统，专为密集运算任务所设计，它结合了 36 个 GB200，等同于采用 72 个 B200 GPU 与 36 个 Grace CPU，可于超大规模的 AI 云计算中实现网络加速、组合存储、零信任安全，以及弹性的 GPU 运算能力。

图 6：台股各散热厂商月度营收及同比情况



资料来源：各公司官网，国海证券研究所

AI 成为全球服务器散热厂商业绩增长的关键动能，预计 2024 年液冷服务器的占比将迅速提升，3D VC 散热方案也有望持续增长。据 chinatimes，奇鋆为 CSP（云端服务商）客户提供 3D VC 散热方案将有所成长，并积极发展下一代液冷技术，目前已为 CSP 客户提供液冷解决方案，预估 2025 年会有更多贡献。法人估计双鸿 Q1 水冷占伺服器营收已达双位数水准，将在下半年显著成长，预计全年水冷营收将占 5% ~ 10%。

表 8: 散热大厂加速布局液冷散热, 东南亚基地产能扩增

公司	具体内容
奇鋳	<ul style="list-style-type: none"> ● 2月营收同比增长, 营收预期两位数增长。2月合并营收 45.53 亿元, 环比-19.8%, 同比+6.9%; 累计前两月营收为 102.29 亿元, 同比+44.7%。 ● 3D VC 方案有所增长, 积极发展液冷技术。据 chinatimes, 奇鋳为 CSP (云端服务商) 客户提供 3D VC 散热方案将有所成长, 并积极发展下一代液冷技术, 目前已为 CSP 客户提供液冷解决方案, 预估 2025 年会有更多贡献。
双鸿	<ul style="list-style-type: none"> ● 2024 年营收或将同比增长 19.81%。2月合并营收 7.35 亿元, 环比-42.5%, 同比-25.7%; 累计 2024 年前 2 月营收为 20.12 亿元, 同比+5.2%。法人估计双鸿 2024 年营收将同比增长近 2 成, 达 19.81%, 获利也将同步成长。 ● 2024Q1 分歧管开始贡献营收, Q3 出货液冷分配器(CDU)。法人估计 Q1 水冷占伺服器营收已达双位数水准, 将在下半年显著成长, 预计全年水冷营收将占 5%~10%。
Vertiv	<ul style="list-style-type: none"> ● 2023Q4, 公司营收 18.65 亿美元, 同比+13%; 调整后营业利润 3.3 亿美元, 同比+57%; 在手订单同比增长 23%, 季度环比增长 18%。2023 年公司营收 68.63 亿美元、同比+21%; 营业利润 10.54 亿美元、同比+140%。 ● 2024 展望: 公司预期营收 75.15-76.55 亿美元、同比+10%; 运营利润 12.75-13.25 亿美元、同比+23.34%。公司表示, 2024 年有望继续保持良好势头, 随着 AI 的数据中心需求推动了额外的市场需求, 公司看到了未来的巨大机遇。 ● 收购: 公司 2023 年第四季度对 CoolTera 实现收购, 与其达成了三年的技术合作伙伴关系。CoolTera 是直接芯片级液冷领导者。 ● 产能扩张: 2024Q4, 公司预期会实现 CoolTera+ 3 Vertir Plants, 预期 2024Q4 季度产能将为 2023Q4 的 45 倍。

资料来源: 科技新报, chinatimes, 台湾经济日报, 优分析, 国海证券研究所

(2) 内地服务器散热: 英维克融合液冷机组首次亮相, 液冷智算中心建设加速

飞荣达、曙光数创、英维克等本土液冷厂商交付良好。飞荣达、曙光数创均表示目前散热产品量产交付、高速增长; 英维克新品频发, 业内首创用于 5MWh 储能系统及工商业储能融合液冷机组将亮相。IDC 预计, 2027 年, 中国液冷服务器市场规模将达 89 亿美元。智算中心建设迎来新趋势, 润泽科技交付了业内首例整栋纯液冷绿色智算中心。

表 9: 液冷产品升级迭代加速, 智算中心建设迎来新趋势

公司	详细内容
飞荣达	<ul style="list-style-type: none"> ● 散热产品量产交付。2月 21 日公司表示, 目前已量产交付的产品包括导热材料、压铸件、风扇、散热模组等。 ● 3月 14 日, 公司表示散热产品包含 VC、热管、风扇、导热材料、石墨片、石墨烯、3D VC 散热器、特种散热器、单相液冷冷板模组, 两相液冷冷板模组等。
曙光数创	<ul style="list-style-type: none"> ● 冷板液冷营收高速增长。公司 2023 全年实现营业收入 6.50 亿元, 同比+25.63%, 其中冷板液冷基础设施产品全年实现收入 1.90 亿元, 同比+430.66%。 ● 液冷服务器占比有望提升。公司表示冷板液冷基础设施产品收入占比由 2023 年同期 6.92% 升至 2024 年的 29.22%, 并表示液冷服务器占目前服务器市场 10% 左右, 随着 AI 和算力需求不断增长, 液冷服务器的占比有望进一步提升。
英维克	<ul style="list-style-type: none"> ● 3月 11-12 日, 2024 数据中心冷却高峰论坛(大湾区)将在广州举办。届时, 面向数据、算力业务的英维克 Coolinside 全链条液冷解决方案也将在论坛上亮相。 ● 3月 11-13 日, 第十三届 CIES 中国国际储能大会将在杭州国际博览中心召开。英维克在业内率先推出的适用于 5MWh 储能系统及工商业储能融合液冷机组将首次公开亮相; 此外, 英维克还将展出 BattCool 储能风冷、液冷机组, 储能专用 SoluKing 长效液冷工质, 快速接头等多款产品。

资料来源: 英维克公众号, iFinD, 国海证券研究所

2.3、交换机：英伟达发布 X800 系列适配 AI 计算，联想/超微等预期 2025 年供货

3月18日，NVIDIA 发布专为大规模 AI 量身定制的全新网络交换机 - X800 系列。NVIDIA Quantum-X800 InfiniBand 网络和 NVIDIA Spectrum™-X800 以太网网络是全球首批高达 800Gb/s 端到端吞吐量的网络平台，将计算和 AI 工作负载的网络性能提升到了新的水平，与其配套软件强强联手可进一步加速各种数据中心中的 AI、云、数据处理和高性能计算 (HPC) 应用，包括基于最新的 NVIDIA Blackwell 架构产品的数据中心。

合作伙伴预期 2025 年开始提供 X800 系列网络平台。全球多家头部基础设施供应商和系统厂商将在明年开始提供基于 Quantum-X800 和 Spectrum-X800 的网络平台，包括 Aivres、DDN、戴尔科技、Eviden、Hitachi Vantara、慧与、联想、超微和 VAST Data 等。

表 10: 英伟达发布 X800 系列交换机，并提供全套软件方案

公司	具体内容
Quantum-X800 平台	<ul style="list-style-type: none"> ● 树立了 AI 专用基础设施极致性能的新标杆，该平台包含了 NVIDIA Quantum Q3400 交换机和 NVIDIA ConnectX®-8 SuperNIC，二者互连达到了业界领先的端到端 800Gb/s 吞吐量，交换带宽容量较上一代产品提高了 5 倍，网络计算能力更是凭借 NVIDIA 的 SHARP™ 技术 (SHARPV4) 提高了 9 倍，达到了 14.4Tflops
Spectrum-X800 平台	<ul style="list-style-type: none"> ● 为 AI 云和企业级基础设施带来优化的网络性能。借助 800Gb/s 的 Spectrum SN5600 交换机和 NVIDIA BlueField-3 SuperNIC，Spectrum-X800 平台为多租户生成式 AI 云和大型企业级用户提供各种至关重要的先进功能。 ● Spectrum-X800 通过优化网络性能，加快 AI 工作负载的处理、分析和执行速度，进而缩短 AI 解决方案的开发、部署和上市时间。Spectrum-X800 专为多租户环境打造，实现了每个租户的 AI 工作负载的性能隔离，使业务性能能够持续保持在最佳状态，提升客户满意度和服务质量。
全套软件方案	<ul style="list-style-type: none"> ● 提供面向万亿参数级 AI 模型性能优化的网络加速通信库、软件开发套件和管理软件等全套软件方案 ● 其中的 NVIDIA 集合通信库 (NCCL) 可将 GPU 的并行计算任务扩展到 Quantum-X800 网络，利用其基于 SHARPV4 的强大网络计算能力和对 FP8 的支持，为大模型训练和生成式 AI 提供超强的性能。 ● NVIDIA 的全栈软件方案带来了先进的可编程性，使数据中心网络变得更加灵活、可靠和灵敏，既提高了整体运营效率，又满足了现代应用和服务的需求。

资料来源：iFinD，国海证券研究所

2.4、光模块：1.6T 研发加速，市场空间广阔

1.6T 产品研发加速，光模块厂商持续扩产。根据 Yole 的数据，2022 年到 2028 年硅光子芯片的市场规模的 CAGR 达到 44%，其中数通光模块在硅光子芯片市场中占比达到 90%以上，为主要应用场景。**新易盛表示加速硅光、相干光模块、1.6T 光模块等行业前沿领域研究及商用；天孚通讯表示持续信息系统升级和自动化升级、江西生产基地降本增效，产能利用率明显提升。**

表 11：光模块厂商持续扩产，1.6T 产品研发加速

公司	具体内容
中际旭创	<ul style="list-style-type: none"> ● 2023 年营收高速增长。2023 年，公司营业收入 1,072,479.41 万元，同比+11.23%；归属于上市公司股东的净利润 218,098.39 万元，同比+78.19%。 ● 800G: 3 月 10 日公司表示，受益于 800G 等高端产品出货比重显著增加及产品设计不断优化，2023 年净利润增速高于营收增速。
新易盛	<ul style="list-style-type: none"> ● 3 月 5 日，公司表示是国内少数批量出货并交付运用于数据中心市场的 400G、800G 高速光模块、掌握高速率光器件芯片封装和光器件封装的企业。公司一向重视行业新技术、新产品的研究，目前已成功推出基于硅光解决方案的 400G、800G 光模块产品及 400G ZR/ZR+相干光模块产品、以及基于 LPO 方案的 800G 光模块产品。 ● 3 月 5 日，公司表示力争抓住 AI 及云数据中心等核心领域良好的市场契机，聚集优势资源持续提升高速率光模块市场占有率，加速硅光、相干光模块、1.6T 光模块等行业前沿领域研究及商用。
天孚通讯	<ul style="list-style-type: none"> ● 3 月 11 日，公司发布 2023 年业绩快报，2023 年公司营收 19.39 亿元、同比+62.07%；归母净利润 7.30 亿元、同比+81.14%。公司实现收入、利润增长，这主要得益于公司前瞻布局和研发投入的增加；公司持续信息系统升级和自动化升级、江西生产基地降本增效，产能利用率明显提升。
剑桥科技	<ul style="list-style-type: none"> ● 2023 年，公司实现营业收入 30.87 亿元，同比-18.46%；归母净利润 0.95 亿元、同比-44.59%。 ● 2023 年，光电子事业部成功完成了新一代 800G 8×FR1 硅光产品和 800G 2×FR4 硅光产品的开发工作，并启动了客户认证测试流程。公司加大了 1.6T 光模块产品的研发投入，计划在 2024 年 OFC（美国光网络与通信研讨会及博览会）中展示研发样机。
光迅科技	<ul style="list-style-type: none"> ● 公司高端光电子器件产业基地将于 2024Q2 逐步完成搬迁，一期项目设计产能可达 100 亿元；具备柔性生产交付能力，未来将会做好产能规划，以期更好地满足客户需求。3 月 12 日，公司表示一直加大海外客户的拓展力度。 ● 市场份额: 2 月 29 日公司表示，据 Omdia 报告，公司最新的全球市场份额位列第四。 ● 800G/1.6T: 2 月 29 日公司表示，800G 和 1.6T 的光模块已经部分应用了自研光芯片。

资料来源：iFinD，国海证券研究所

2.5、主板：厂商产能持续爬坡，2024 年预期增长

主板厂商产能持续爬坡，2024 年预期恢复增长。深南电路表示，无锡基板二期工厂、广州封装基板项目一期工厂已先后连线，目前均处于产能爬坡阶段。华擎表示，预期今年 AI 伺服器将加入贡献，将推升伺服器业务较去年明显增长。技嘉表示，主板出货动能回温，2024Q1 预期显卡在新一代 GPU 新品陆续推出下将持续带动出货。

表 12: 基板厂商产能持续爬坡, 2024 年预期恢复增长

公司	具体内容
技嘉	<ul style="list-style-type: none"> ● 2024 年 2 月, 营收 139.69 亿元, 环比-17.5%, 同比+38.5%, 达历年同期新高; 累计前两月营收为 308.99 亿元, 同比+68.8%。受惠 CPU 新世代新品效益及传统小旺季, 技嘉主板出货动能回温。展望 2024Q1, 显卡在新一代 GPU 新品陆续推出下将持续带动出货。
华擎	<ul style="list-style-type: none"> ● 2024 年 2 月, 营收 13.77 亿新台币, 同比+3.92%; 累计前两月营收为 33.47 亿新台币, 同比+21.39%。公司表示伺服器方面, 预期今年 AI 伺服器将加入贡献, 将推升伺服器业务较去年明显增长, 整体而言, 华擎今年营运将较去年升温。
沪电股份	<ul style="list-style-type: none"> ● 3 月 6 日, 公司表示泰国工厂目前规划主要是应用于通信通讯、数据中心、汽车电子等领域的印制电路板, 产品主用应用于通信通讯设备、包括 AI 在内的数据中心基础设施、汽车电子等领域。
深南电路	<ul style="list-style-type: none"> ● 3 月 15 日, 公司发布 2023 年年报, 公司营业收入 135.26 亿元, 同比-3.33%; 归母净利润 13.98 亿元, 同比-14.81%。上述变动主要由于下游市场需求下行, 封装基板和 PCB 业务全年整体稼动率较上年同期有所下降, 叠加封装基板新项目建设、新工厂爬坡等带来的费用和固定成本增加等因素影响。 ● 2 月 22 日, 公司表示高阶倒装芯片用 IC 载板产品制造项目(无锡基板二期)已于 2022 年 9 月下旬连线投产, 目前处于产能爬坡阶段。公司拥有印制电路板、电子装联和封装基板三项主营业务, 下游应用领域广泛, 覆盖通信、数据中心、工控医疗、汽车电子等领域。 ● 2 月 5 日, 公司表示无锡基板二期工厂、广州封装基板项目一期工厂已先后连线, 目前均处于产能爬坡阶段。不同工厂面向的产品类型不同, 实际达成产能与市场环境、产品结构、技术改造进程等因素相关。

资料来源: iFinD, 国海证券研究所

3、GPU 产业链: 台积电 3nm/4nm 产能满载, Meta 新建集群含超 4.9 万块 H100

3.1、COWOS: 台积电 3nm/4nm 产能满载

台积电订单需求强劲, 产能大幅增长。AI 芯片带动先进封装, 英伟达下一代芯片将采用台积电 3/4nm 制程, 强劲订单推动先进制程产能满载、营收淡季不淡。据 3 月 13 日财联社, 台积电 3 月对 CoWoS 设备厂再次启动新一波追单, 交机时间预计为今年四季度。原本预计 2024 年底台积电 CoWoS 月产能将达到 3.2 万~3.5 万片, 如今预期或超过 4 万片。

表 13: 台积电 3nm、4nm 产能近满载，预计 Q4 CoWoS 月产能超 4 万片

分类	具体内容
产能	<ul style="list-style-type: none"> ● 据 3 月 13 日财联社，台积电 3 月对 CoWoS 设备厂再次启动新一波追单，交机时间预计为今年四季度。原本预计 2024 年底台积电 CoWoS 月产能将达到 3.2 万~3.5 万片，如今预期或超过 4 万片。 ● 据 3 月 20 日财联社，2024 年英伟达对 CoWoS 需求将较去年成长三倍。对此，黄仁勋未正面回应相关数字，不过，他两次强调今年对于 CoWoS 的需求非常高。
制程	<ul style="list-style-type: none"> ● 2023Q4，台积电独占晶圆代工行业 61.2% 营收，环比+14%，达 196.6 亿美元。其中，先进制程 7nm（含）以下营收比重上升至 67%；伴随 3nm 产能与投片逐季到位，先进制程营收比重有望突破七成大关。 ● 台积电 3nm 已拿下苹果、高通及联发科等大厂订单，业界预期，台积电今年将全力扩增 3nm 产能，甚至将调配部分 5nm 产能转至 3nm，预计今年底前台积电 3nm 产能利用率有望突破 80%。英伟达 H200 及新一代 B100 将分别采台积电 4nm 及 3nm 制程。法人指出，英伟达订单强劲，台积电 3nm、4nm 产能几近满载，首季运营淡季不淡。
扩产	<ul style="list-style-type: none"> ● 据 3 月 18 日财联社，台积电将在嘉义科学园区先进封装厂新厂加大投资，园区将拨出六座新厂用地给台积电，比原本预期的四座多两座，总投资额逾 5000 亿台币，主要扩充 CoWoS 先进封装产能，相关环评、水电设施都已盘点、处理完成，预计 4 月上旬对外公布。 ● 为应对客户对 AI 芯片需求，台积电再度启动大扩产，预计四月将展开装机与在建中的 2nm 晶圆厂、先进封装厂共达十座。另外，供应链指出，继 2nm 之后，台积电还将推进到 1.4nm、1nm 先进制程；厂商透露，1nm 世代在台湾地区投资建厂大概需要 8-10 座厂。

资料来源：财联社，爱集微，《科创板日报》，国海证券研究所

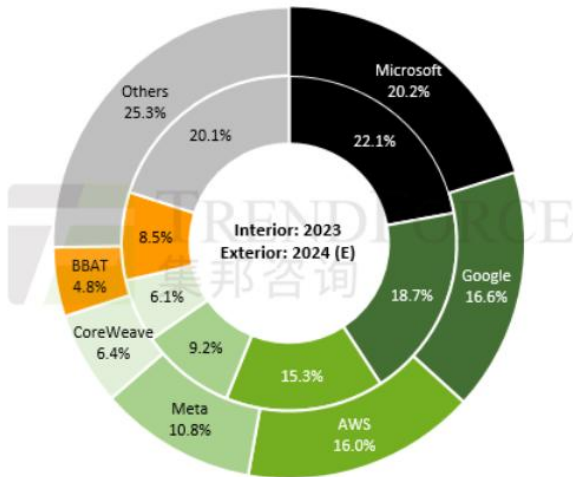
3.2、互联网需求旺盛：Meta 新建集群含超 4.9 万块 H100

CSP 的 AI 服务器需求量全球领先。根据 TrendForce 集邦咨询最新预估，以 2024 年全球主要云端服务业者（CSP）对高端 AI 服务器需求量观察，预估美系四大 CSP 业者包括 Microsoft、Google、AWS、Meta 各家占全球需求比重分别达 20.2%、16.6%、16% 及 10.8%，合计将超 6 成。其中，又以搭载 NVIDIA GPU 的 AI 服务器机种占大宗。

Meta 全年 H100 需求乐观。根据 3 月 13 日 IT 之家，Meta 公布了两座新的数据中心集群，内含超 4.9 万块英伟达 H100 GPU，专门训练 Llama3。Meta 表示 2024 年年底，将拥有大约 35 万片英伟达 H100 加速卡，如果算上其它 GPU 的话，其计算能力相当于 60 万片 H100。

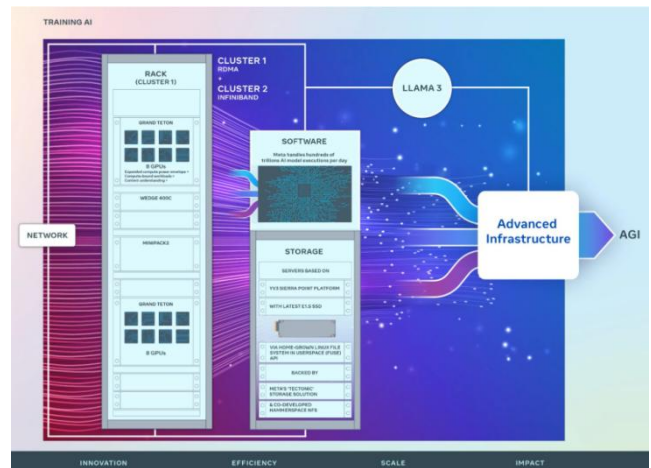
MI300 或低价出售给 CSP 厂商。根据 2 月 14 日 IT 之家，AMD 为 Instinct MI300 系列加速器谋发展，将以较低的价格供货给微软公司，促进双方新一轮长期合作。Meta、OpenAI 和微软在 AMD 投资者活动上表示，他们都将使用 AMD 最新开发的人工智能芯片 Instinct MI300X。

图 7：2023~2024 年全球 CSP 高阶 AI 服务器需求



资料来源：TrendForce

图 8：Meta 新建两座数据中心集群



资料来源：IT之家

4、相关公司

行业评级及投资策略：大模型带动 AI 算力需求增长，算力产业链中的 AI 芯片厂商与下游服务器整机厂商有望持续受益。维持对计算机行业“推荐”评级。

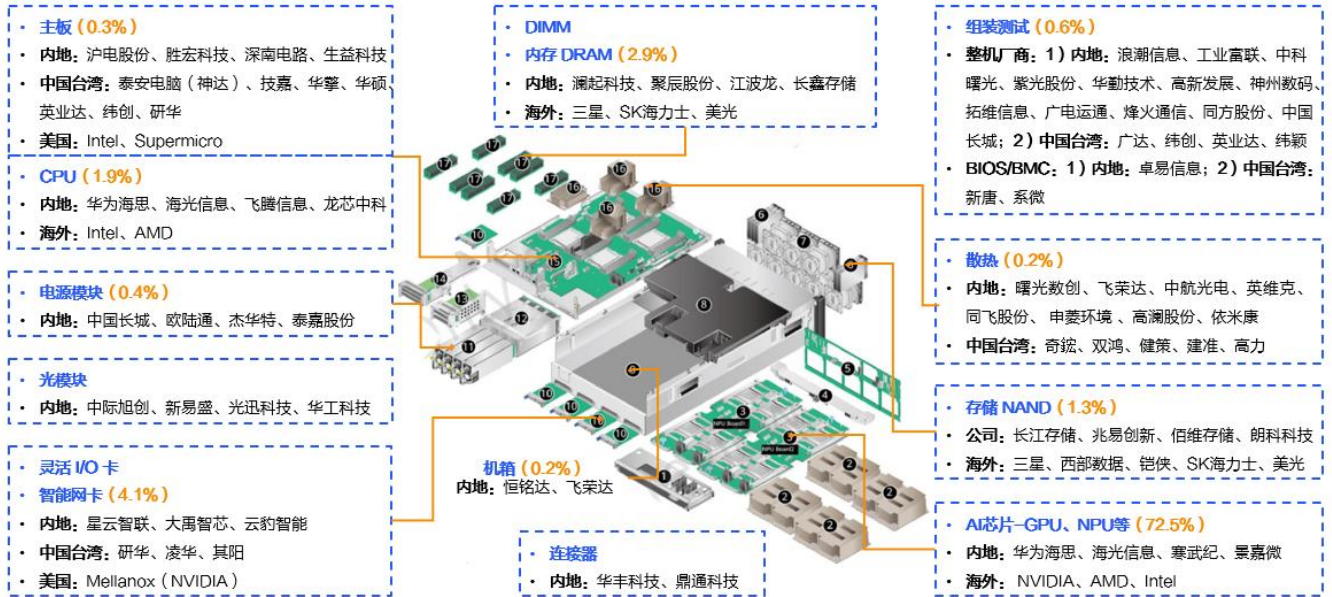
1) 服务器整机：工业富联、浪潮信息、中科曙光、华勤技术、中国长城、高新发展、神州数码、烽火通信、拓维信息、纬创、广达、英业达、纬颖、超微电脑。

2) 服务器组件：
①AI 芯片：海光信息、寒武纪、龙芯中科、景嘉微、英伟达、AMD、Intel；
②散热：飞荣达、曙光数创、英维克、同方股份、申菱环境、高澜股份、奇鋆科技、双鸿、VERTIV；
③主板：沪电股份、深南电路、胜宏科技、技嘉、华擎。

3) 光模块：天孚通信、中际旭创、新易盛、光迅科技、华工科技。

4) 数据中心：奥飞数据、光环新网、宝信软件、数据港、电科数字。

图 9：服务器产业链相关公司



注: 蓝字为零部件, 括号中橙色数字为价值量。

上图展示以华为 Atlas 800 训练服务器为例 (鲲鹏 920 * 4, 昇腾 910 * 8), 其中 CPU 集成在主板上, NPU 集成在 NPU 板上; 具体价值量数字对标 Nvidia DGX H100 服务器, 具体计算方法见下文

资料来源: 华为昇腾官网, Semianalysis, 各公司公告, 各公司官网, 百度百科, 爱采购, 国海证券研究所

5、风险提示

- 1) 宏观经济影响下游需求:** 宏观经济环境下行, 将影响客户对信息化基础设施的采购需求;
- 2) 信创政策不及预期:** 行业主要驱动因素之一是信创政策持续落地, 若信创产业推进不及预期, 或导致行业内公司订单增速下行;
- 3) 市场竞争加剧:** IT 产品和服务行业是成熟且完全竞争的行业, 新进入者可能加剧整个行业的竞争态势;
- 4) 中美博弈加剧:** 国际形势持续不明朗, 美国不断通过“实体清单”等方式对中国企业实施打压, 若中美紧张形势进一步升级, 将可能导致中国半导体供应链供应受到影响;
- 5) 相关公司业绩不及预期:** 市场环境变化、公司治理情况变化、其他非主营业务经营不及预期等原因或将导致相关公司的整体业绩不及预期。
- 6) 各公司并不具备完全可比性, 对标的相关资料和数据仅供参考。**

【计算机小组介绍】

刘熹，计算机行业首席分析师，上海交通大学硕士，多年计算机行业研究经验，致力于做前瞻性深度研究，挖掘投资机会。新浪金麒麟新锐分析师、Wind 金牌分析师团队核心成员。

【分析师承诺】

刘熹，本报告中的分析师均具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观的出具本报告。本报告清晰准确的反映了分析师本人的研究观点。分析师本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收取到任何形式的补偿。

【国海证券投资评级标准】

行业投资评级

推荐：行业基本面向好，行业指数领先沪深 300 指数；
中性：行业基本面稳定，行业指数跟随沪深 300 指数；
回避：行业基本面向淡，行业指数落后沪深 300 指数。

股票投资评级

买入：相对沪深 300 指数涨幅 20%以上；
增持：相对沪深 300 指数涨幅介于 10%~20%之间；
中性：相对沪深 300 指数涨幅介于-10%~10%之间；
卖出：相对沪深 300 指数跌幅 10%以上。

【免责声明】

本报告的风险等级定级为 R3，仅供符合国海证券股份有限公司（简称“本公司”）投资者适当性管理要求的客户（简称“客户”）使用。本公司不会因接收人收到本报告而视其为客户。客户及/或投资者应当认识到有关本报告的短信提示、电话推荐等只是研究观点的简要沟通，需以本公司的完整报告为准，本公司接受客户的后续问询。

本公司具有中国证监会许可的证券投资咨询业务资格。本报告中的信息均来源于公开资料及合法获得的相关内部外部报告资料，本公司对这些信息的准确性及完整性不作任何保证，不保证其中的信息已做最新变更，也不保证相关的建议不会发生任何变更。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，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波动。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。报告中的内容和意见仅供参考，在任何情况下，本报告中所表达的意见并不构成对所述证券买卖的出价和征价。本公司及其本公司员工对使用本报告及其内容所引发的任何直接或间接损失概不负责。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或者金融产品等服务。本公司在知晓范围内依法合规地履行披露义务。

【风险提示】

市场有风险，投资需谨慎。投资者不应将本报告视为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向本公司或其他专业人士咨询并谨慎决策。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。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息。本报告不构成本公司向该机构之客户提供的投资建议。

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。本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

【郑重声明】

本报告版权归国海证券所有。未经本公司的明确书面特别授权或协议约定，除法律规定的情况外，任何人不得对本报告的任何内容进行发布、复制、编辑、改编、转载、播放、展示或以其他方式非法使用本报告的部分或者全部内容，否则均构成对本公司版权的侵害，本公司有权依法追究其法律责任。