

AI大模型风起云涌，半导体与光模块长期受益

半导体行业深度报告（十）

2024年3月14日

证券分析师：方霁，执业证书编号：S0630523060001

联系人：蔡望颀

联系方式：cwt@longone.com.cn



摘要

- **全球AI大模型高速发展，算力需求高增驱动AI服务器三年CAGR约为29%的增长，带动算力芯片与光模块产业链受益。**2023年以来，以ChatGPT、Sora为代表的多模态AI大模型横空出世，标志着人工智能技术已经进入一个新的纪元。未来，通用人工智能（AGI）有望集多模态感知、大数据分析、机器学习、自动化决策于一体，重塑人类工作和生产生活方式，引领人类步入第四次工业革命。算力的高速增长需要更多的AI服务器支撑，2023年全球AI服务器约85.5万台，到2026年预计将达到236.9万台，CAGR为29.02%，从而驱动AI算力芯片与配套的光模块产业高增长。
- **GPU是常见的AI芯片种类，AI芯片一般占据AI服务器成本70%左右，国产算力芯片在海外垄断格局下有望实现国产替代。**AI芯片按照技术架构和应用需求可分为GPU、FPGA、ASIC和类脑芯片四大类，GPU是多功能的并行处理器，由于其通用程度高、软件生态丰富、制造工艺相对成熟，是目前最为普遍的AI芯片类型，占到中国AI运算市场的约89%。GPU是AI服务器的核心，约占近90%AI芯片市场份额，其价值量占AI服务器高达70-75%。2023Q4英伟达、AMD、英特尔分别占据全球GPU市场份额是80%、19%、1%。中国AI算力在文心一言、讯飞星火、通义千问等大模型支持下，长期需求规模较大。
- **HBM一定程度解决了算力增速大于存储增速的内存墙问题，由于其极高带宽、低功耗、小体积优势，成为GPU显存的最佳方案，随着AI算力芯片的高增长，HBM飞快发展，国内相关产业链企业或将受益。**近几十年来，处理器的性能以每年大约55%速度快速提升，而内存性能的提升速度则只有每年10%左右。不均衡的发展速度造成了当前内存的存取速度严重滞后于处理器的计算速度，内存瓶颈导致高性能处理器难以发挥出应有的功效。HighBandwidth Memory，即高带宽内存，是一种新兴的DRAM 解决方案。HBM具备极高带宽：达到1T/s；体积减小：比GDDR降低94%的尺寸；低功耗：高度集成后比GDDR拥有更小的电压与功耗。这些显著优势促使HBM快速发展，目前全球主要被韩美企业垄断，国内厂商纷纷布局，适合国产HBM发展的产品即将问世。
- **光模块受益算力需求高增长，我国模块封装能力成熟，或将受益高速光模块需求增长，同时受益光芯片国产化进程。**光模块用于服务器或者数据中心的高速互联，主要下游在电信与IDC，随着AI服务器发力发展，数通光模块在2026年或将占据60%份额。全球TOP10大光模块企业中，中国大陆占据5家，封装能力全球领先。从光模块成本结构看，光模块器件占据了光模块73%的成本，而光芯片与电芯片占据光器件的主要成本，高端光芯片（25G以上）国产替代率较低，国内企业在2.5G和10G光芯片领域基本实现了核心技术的掌握，国产化率分别为90%和60%，但是25G光芯片国产化率为20%，25G以上光芯片国产化率仅为5%，国产替代空间较大。
- **投资建议：**AI大模型时代下，AI算力需求高速增长，AI服务器需求呈现29%复合高增速，从而驱动算力芯片及光模块的需求高增长。短期关注相关产业链的主题催化行情，长期关注受益于AI持续高速发展业绩或将逐步兑现优质企业。建议关注海光信息、寒武纪、澜起科技、中际旭创、光迅科技、天孚通信、新易盛、源杰科技等优质算力芯片与光器件相关企业。
- **风险提示：**AI需求不及预期风险，行业竞争过度风险，国际贸易政策变化风险。

目录

第一部分 大模型带动AI服务器高增长

第二部分 算力芯片与光模块长期受益

第三部分 A股上市公司代表

第四部分 投资建议

第五部分 风险提示



1.1、通用AI概念：AI有望引领人类第四次工业革命

- ▶ 复盘历史上三次工业革命，每一轮都伴随着核心技术的突破和生产方式的重大变革。第一次工业革命以蒸汽机的发明为代表，机器解放了人类的双手，第二次则由电力和内燃机驱动，改变了人类交通和通信的方式，第三次是计算机和互联网技术的发明，使自动化产线和工业机器人得以大规模应用，移动通讯技术发展使信息传播速度前所未有，极大促进了生产力的发展。
- ▶ 2023年以来，以ChatGPT、Sora为代表的多模态AI大模型横空出世，标志着人工智能技术已经进入一个新的纪元。未来，通用人工智能（AGI）有望集多模态感知、大数据分析、机器学习、自动化决策于一体，重塑人类工作和生产生活方式，引领人类步入第四次工业革命。

第一次工业革命
1760-1840



人工劳动→机械化

核心驱动：蒸汽机的发明和应用使得机器代替了人工，纺织机械的改进加速纺织工业机械化，铁路运输兴起改善了交通和物流。

第二次工业革命
1870-1914



机械化→电气化

核心驱动：内燃机诞生促进了飞机和汽车的发明，交通运输效率大幅提高，化工和钢铁产业快速发展，电话和无线电的发明极大加快了信息传播速度。

第三次工业革命
1960年至今



电气化→数字化

核心驱动：互联网和计算机的诞生极大地提升了生产效率，移动通信和智能手机使通信和信息传播速度大幅提高，自动化生产线和机器人技术在制造业中的应用。

第四次工业革命?
2023-?



数字化→智能化

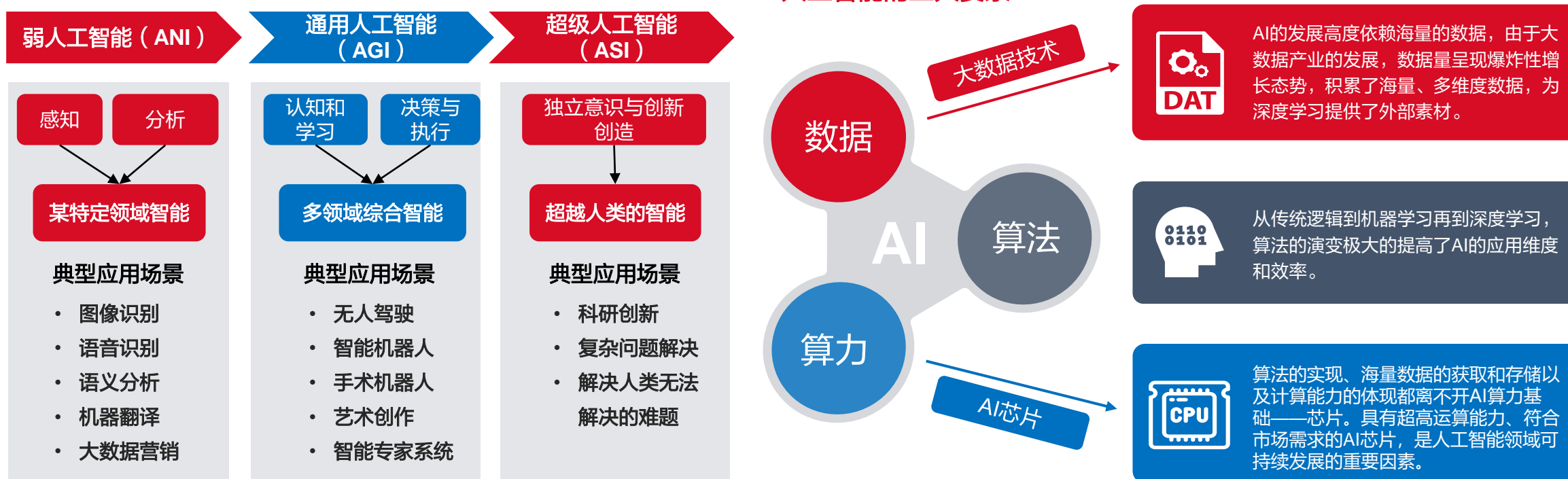
核心驱动：AI技术快速发展，物联网（IoT）实现万物互联，大数据技术成熟促进数据驱动的决策模式，自动驾驶、工业机器人乃至通用人工智能的发展有望重塑人类生产生活方式。

资料来源：公开资料整理，东海证券研究所

1.1、通用AI概念：人工智能的分类和三大要素

- ▶ 人工智能（Artificial Intelligence）是通过计算机和算法来模拟、扩展人类智能的一门技术科学。其本质是使计算机和人一样具备学习、推理、感知和决策的能力，代替人类解决和处理各类复杂的工作，从而提升效率和解放生产力。常见的AI研究包括机器学习、机器视觉、自然语言处理和专家系统等。
- ▶ 按照智能程度划分，AI可分为弱人工智能（ANI）、通用人工智能（AGI）和超级人工智能（ASI）。弱人工智能是指只能解决单个特定领域问题的AI，如面部识别、语音识别等，目前已广泛应用。通用人工智能是指具备人类级别智能的AI，目前还尚未实现，但Sora的问世无疑使我们离AGI更进了一步。超级人工智能是指超越人类智能且具有自主思维意识的AI，目前尚处理论阶段。
- ▶ 人工智能具有算力、算法、数据三大要素，其中基础层提供算力支持，通用技术平台解决算法问题，场景化应用挖掘数据价值。

人工智能的三大要素



资料来源：行行查，公开资料整理，东海证券研究所

1.2、AIGC产业链：基础设施是AI算力之源，下游应用前景广阔

- AIGC 即 AI Generated Content，即利用AI技术自动创建文本、图像、视频等内容，它被认为是继PGC、UGC之后的新型内容创作方式。
- AIGC产业可分为基础设施层、模型层和应用层，每一层都是AIGC产业链不可或缺的组成部分，共同构成了一个完整的生态系统，以支持从数据处理到内容创作的所有环节。

上游：基础设施层：构成AIGC核心的计算和存储平台，包括数据中心、算法平台、以及AI服务器、高性能计算硬件以及云计算服务。

中游：模型层：包括开发和训练各类AI大模型的算法和技术，主要为中美互联网科技巨头如OpenAI、微软、谷歌、百度、阿里等。

下游：应用层：直接面向最终用户的AIGC产品和服务，如C端的多模态生成式AI产品，以及各类B端的垂直行业大模型解决方案。

上游：基础设施层

AI算力



AI算法框架



大数据解决方案



中游：模型层

AI大语言模型



公司名称	预训练模型	时间	参数量 (亿)
Open AI	Sora	2024年2月	30
	GPT-4	2023年3月	17000
	CLIP&DALL-E	2021年1月	120
谷歌	GPT-3	2020年5月	1750
	Gemini	2023年12月	16000
	Bard	2023年3月	1370
Meta	LaMDA	2021年5月	1370
	BERT	2018年10月	3.4
	LLaMA-2	2023年7月	300
百度	OPT-175B	2022年5月	1750
	文心大模型4.0	2023年10月	15000
	文心大模型3.5	2023年6月	2600
	文心大模型3.0	2021年7月	2600

下游：应用层

C端通用生成式AI

- 文本生成** 文案创作、办公辅助、代码生成
- 图像生成** 图像创作、图像修复、虚拟试衣
- 音频生成** 音乐创作、语音合成、聊天机器人
- 视频生成** 影视创作、数字分身、娱乐游戏

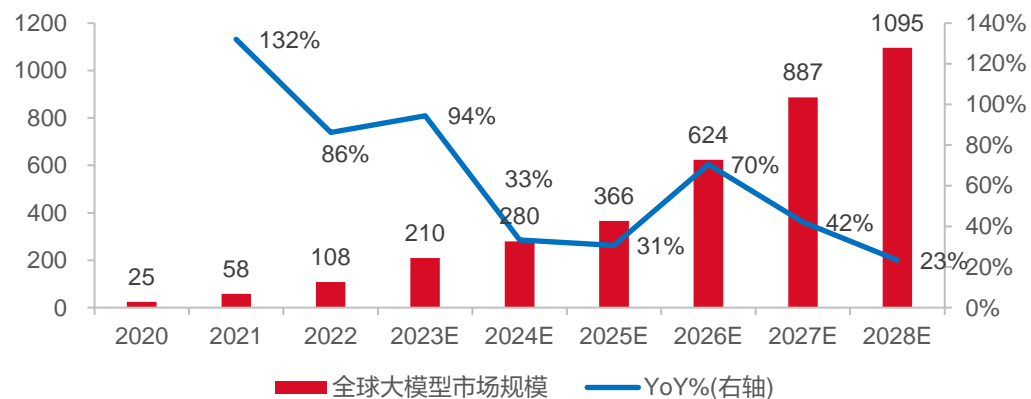
B端垂直大模型解决方案

- 无人驾驶
- 无人工厂
- 影视创作
- 财务法律
- 科学研究
- 投资分析
- 教育学习
- 医疗诊断
-

资料来源：公开资料整理，东海证券研究所

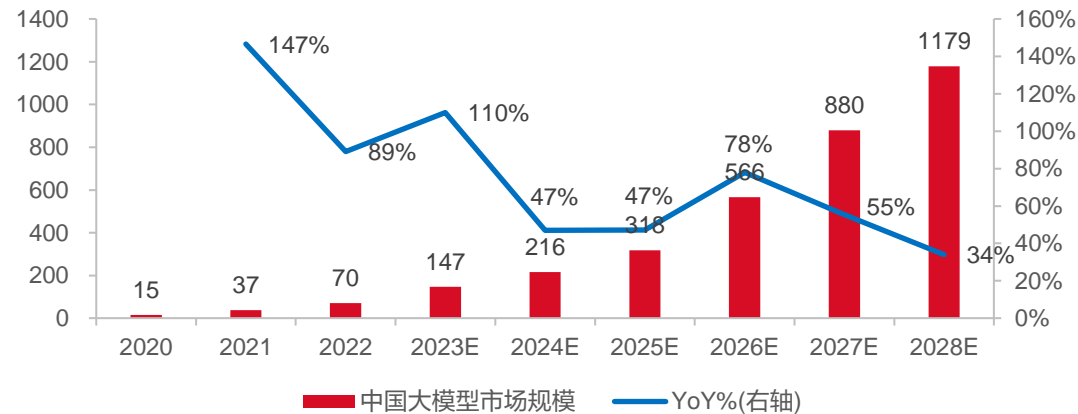
1.3、大模型市场规模：千亿级宽阔赛道，AIGC市场潜力无穷

2020-2028E全球大模型市场规模及预测（亿美元）



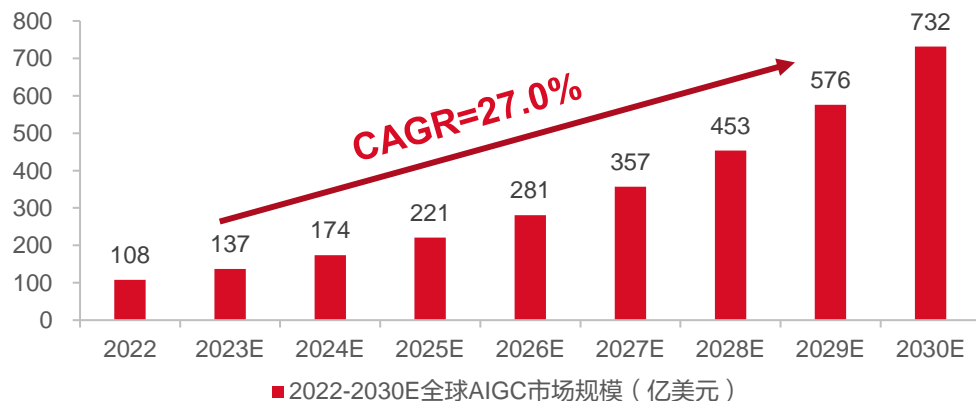
资料来源：大模型之家，东海证券研究所

2020-2028E中国大模型市场规模及预测（亿元）



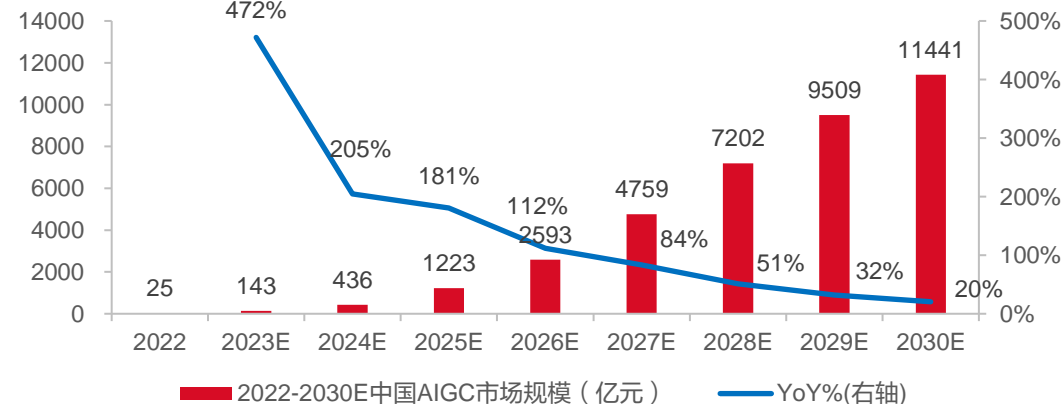
资料来源：大模型之家，东海证券研究所

2022-2030E全球AIGC市场规模及预测（亿美元）



资料来源：Precedence Research, IDC, 东海证券研究所

2022-2030E中国AIGC市场规模及预测（亿元）



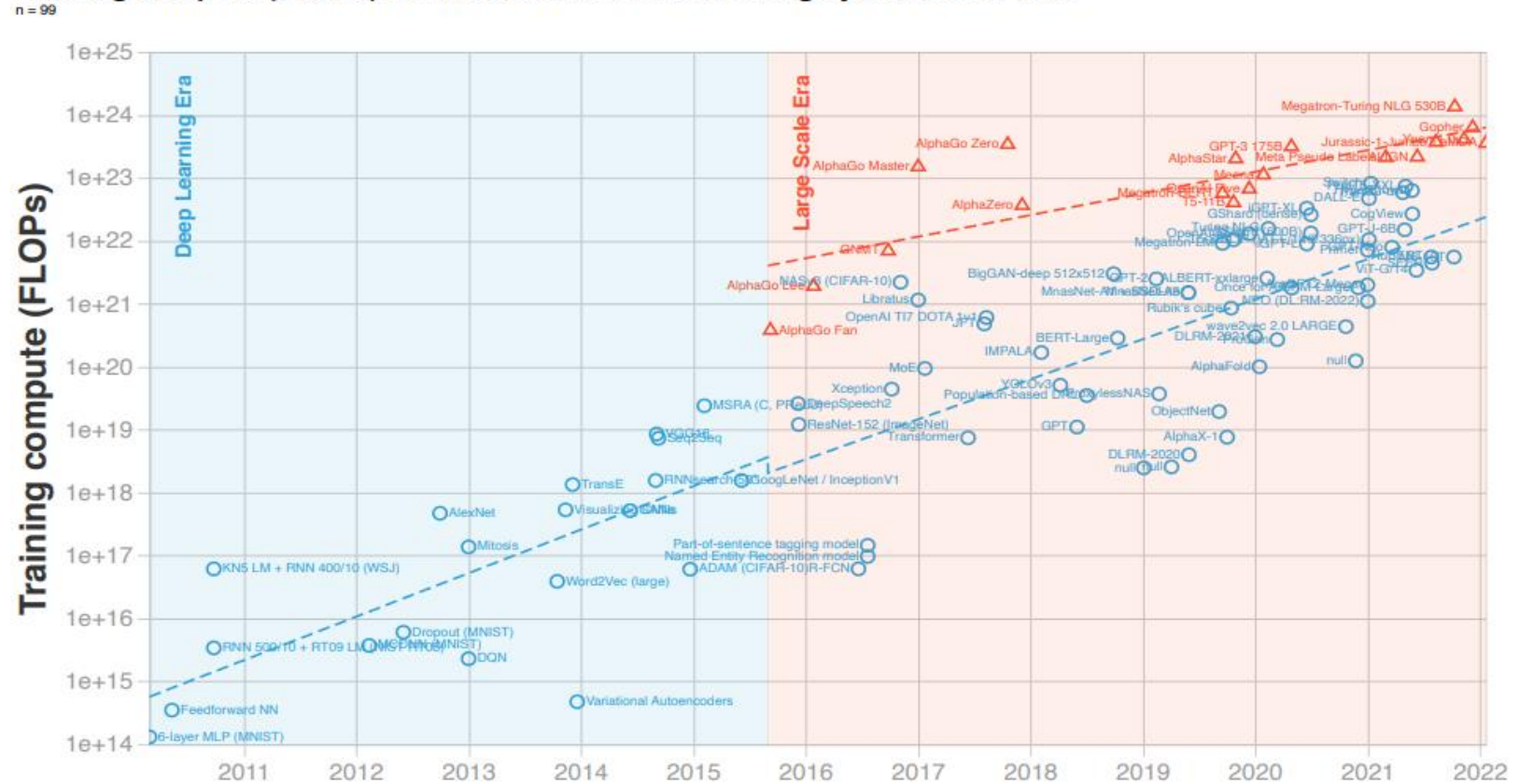
资料来源：艾瑞咨询，东海证券研究所

1.4、算力需求激增：AI大模型的性能和训练算力需求呈显著正相关

- 机器学习的训练计算可分为三个时期：
- 前深度学习时代（1952-2010）：这一时期算力增长主要受CPU和初期GPU的性能提升驱动，训练计算需求大约每20个月翻一番，基本符合摩尔定律。
- 深度学习时代（2010-2015）：随着深度学习技术的兴起，算力需求增速显著加快，GPU开始被大量用于神经网络训练，训练算力翻倍时间缩短至大约5-6个月，超越了摩尔定律。
- 大模型时代（2015-至今）：随着BERT、GPT等千亿乃至万亿级参数规模的大模型涌现，算力需求再次显著增加，尽管算力翻倍时间放缓至10个月左右，但其计算量相较于深度学习时代提升了2-3个数量级。
- 未来，随着ChatGPT、Sora、文心一言等大模型的普及，模型推理所需的算力也会大幅增加，从而进一步提高对AI算力的需求，带动整个AI算力产业链不断增长。

1952-2022年主要机器学习系统的训练算力需求（FLOPs）

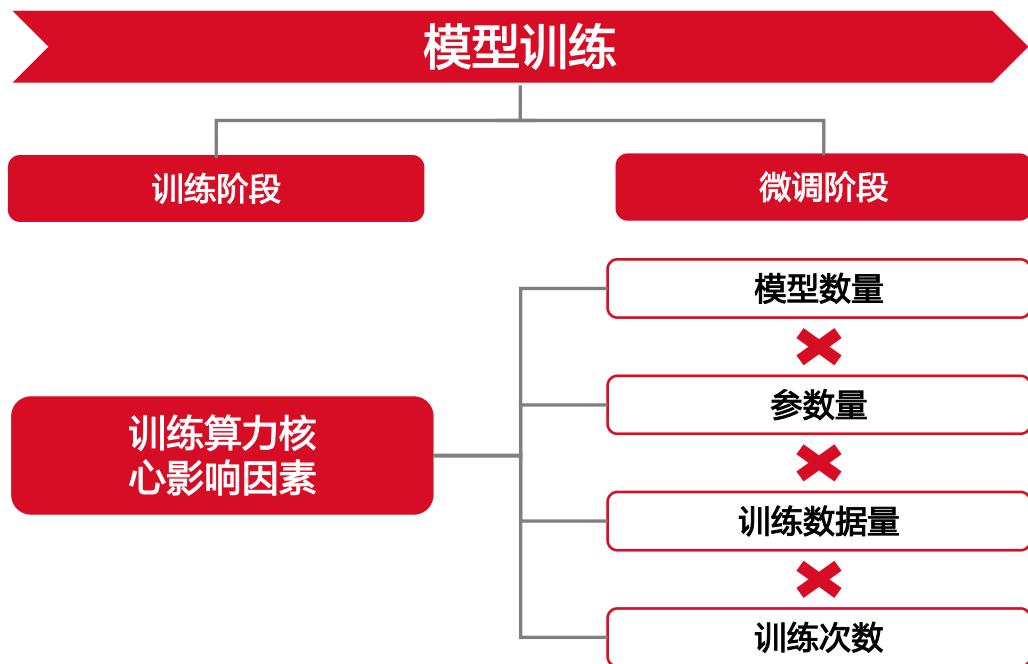
Training compute (FLOPs) of milestone Machine Learning systems over time



资料来源：《Compute Trends across 3 eras of Machine Learning》，J.Sevilla, L.Heim, A. Ho, T.Besiroglu, M.Hobbhahn, P.Villalobos, 东海证券研究所

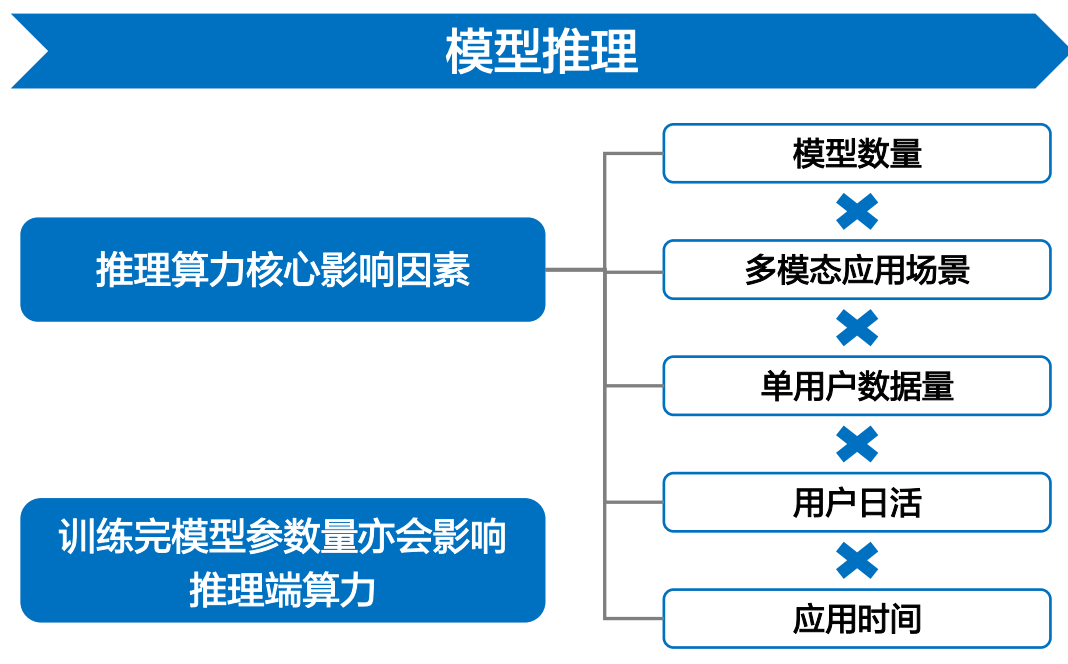
1.4、算力需求激增：AI大模型的训练和推理均离不开强大算力支撑

- AI大模型的实践应用涵盖了两个核心阶段：模型训练和模型推理，这两个环节共同构成了AIGC技术的算力框架。
- **训练**：指通过学习大量数据来不断优化模型参数，以期能够准确响应特定任务，随着参数增加或训练次数增多，其对算力要求亦越高。
- **推理**：则涉及将训练完的模型应用于新数据的输入，以生成有用的内容或决策，其算力需求较训练更低，更侧重处理应用场景中实时数据流的能力，其算力挑战主要来自于用户端响应速度以及对吞吐数量的要求。



如今大模型参数规模已成长至千亿乃至万亿级，训练数据量和训练迭代次数同时增加，推动预训练及对微调阶段算力需求大幅提升。

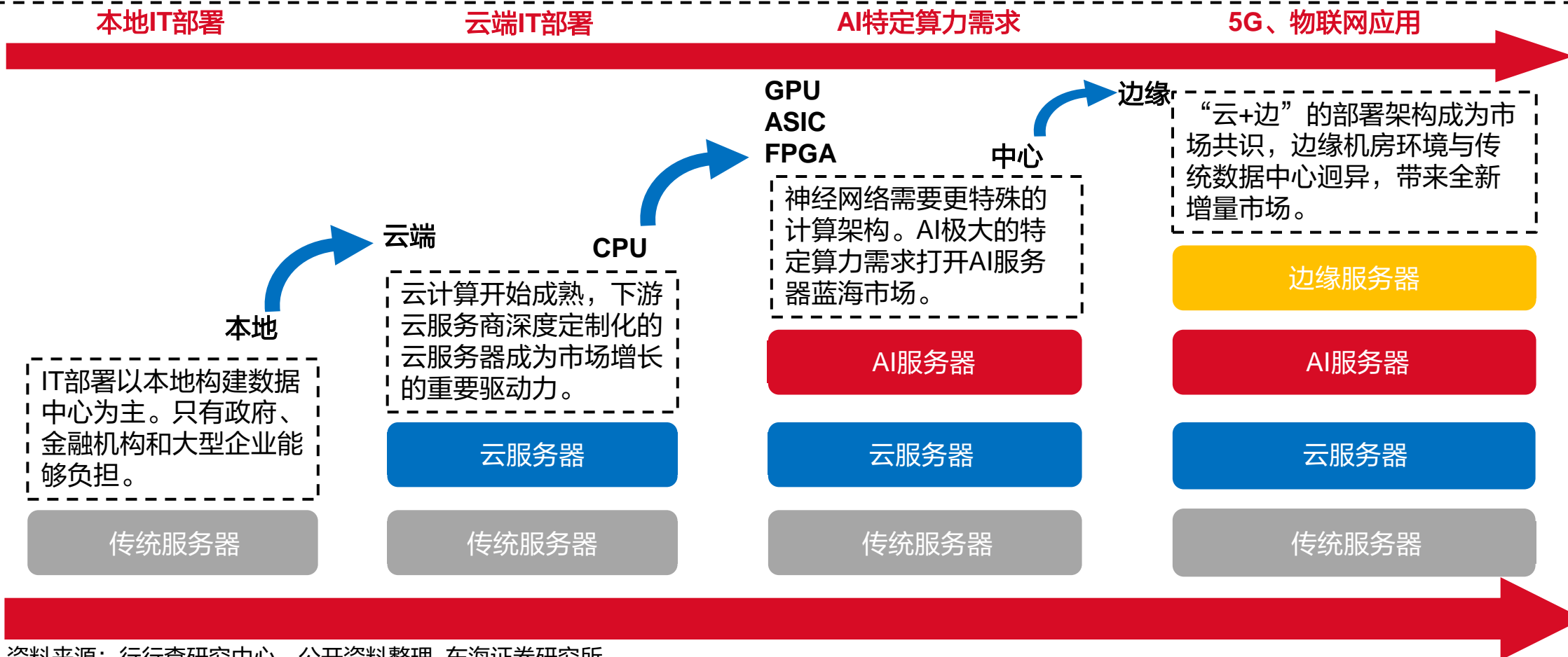
资料来源：甲子光年智库，公开资料整理, 东海证券研究所



多模态应用场景的拓展和用户人数的增长使推理数据处理量和模型部署量快速膨胀，触发了对推理算力的爆炸性增长。

1.5、AI服务器：全球服务器行业的演进历程

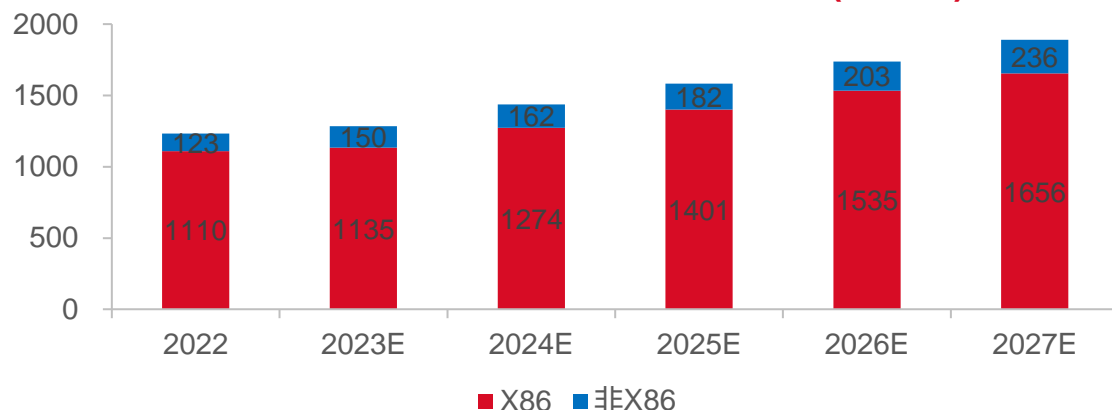
- 2012年，云计算技术的兴起彻底改变了服务器的部署模式，推动企业IT建设从传统的本地环境向云端迁移。
- 2016年，Alpha Go引领的人工智能科技的第三次浪潮催生了对新型架构服务器的迫切需求，为服务器市场注入了新的活力。
- 2020年，5G通信技术的广泛应用为边缘计算领域开辟了广阔的蓝海市场，由于机房环境的迥异，服务器行业迎来全新增量市场。



资料来源：行行查研究中心，公开资料整理, 东海证券研究所

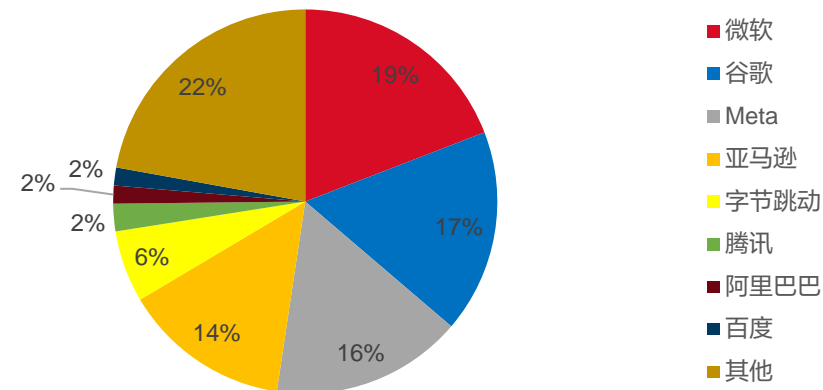
1.5、AI服务器： AI服务器是大模型算力之源，中美科技巨头是主要买家

2022-2027E全球服务器市场规模及预测(亿美元)



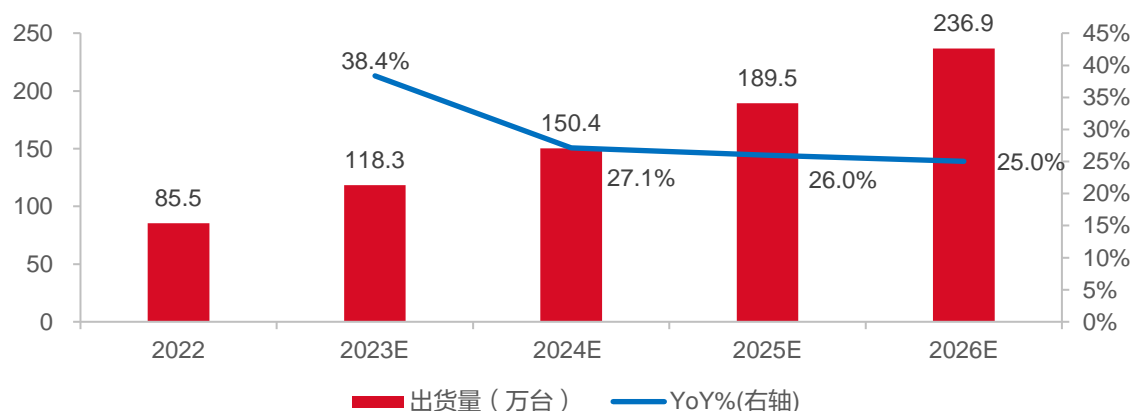
资料来源：IDC, Gigalight, 东海证券研究所

2022年全球AI服务器采购量占比



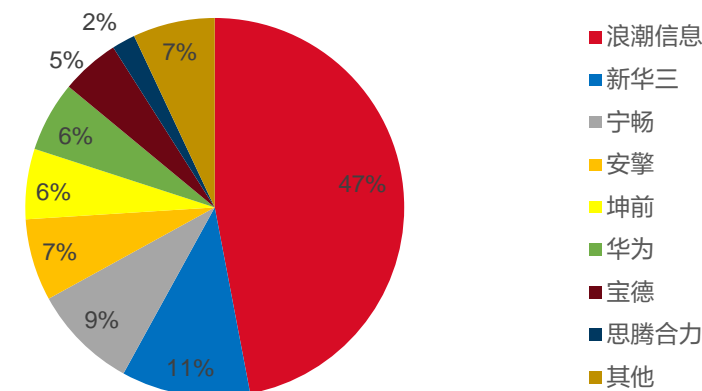
资料来源：TrendForce, 观知海内, 东海证券研究所

2022-2026E全球AI服务器出货量及预测 (万台)



资料来源：观知海内, 东海证券研究所

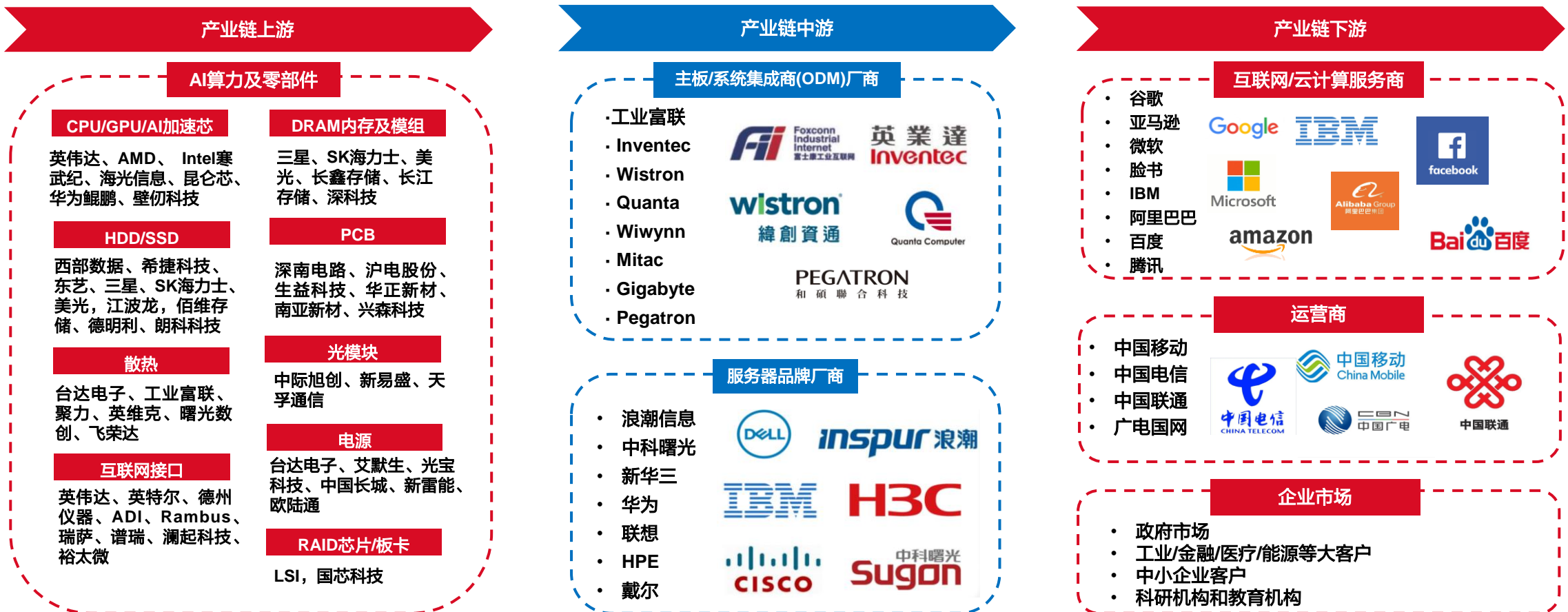
2022中国AI服务器市场份额构成



资料来源：观知海内, 东海证券研究所

1.5、AI服务器： AI服务器产业链解析

- AI服务器产业链的上游厂商主要为电子元件厂商，中游为服务器厂商，下游客户则包括数据中心、政府、各类企业等。
- 核心零部件如算力芯片、DRAM、SSD、RAID芯片市场集中度较高，主要由美、日、韩企业主导，头部厂商市占率仍处于垄断地位，国产厂商整体实力与国外龙头相比尚有差距，但近年来正在加速国产替代步伐。



资料来源：公开资料整理，东海证券研究所

目录

第一部分 大模型带动AI服务器高增长

第二部分 算力芯片与光模块长期受益

第三部分 A股上市公司代表

第四部分 投资建议

第五部分 风险提示



2.1、AI芯片：AI芯片的主要类别和性能对比

- **定义：**AI芯片指面向AI应用，针对AI算法（如深度学习等）进行特殊加速设计的芯片。
- **分类：**根据技术架构和应用需求，AI芯片可分为GPU、FPGA、ASIC和类脑芯片四大类。GPU是多功能的并行处理器，由于其通用程度高、软件生态丰富、制造工艺相对成熟，是目前最为普遍的AI芯片类型，占到中国AI运算市场的约89%。FPGA芯片是可编程的芯片，允许开发者按需定制硬件，在需要特定算法优化时非常有用，可根据算法迭代调整硬件配置。ASIC是为特定AI应用定制的，能在性能和能效上提供最佳的表现，该类芯片是固定设计，针对一种特定任务或算法进行了优化。类脑芯片颠覆传统冯诺依曼架构，是一种模拟人脑神经元结构的芯片，目前尚处于起步阶段。
- **AI芯片（GPU/FPGA/ASIC）在云端兼顾执行人工智能的“训练”与“推理”任务，而在终端主要负责执行“推理”操作。**就性能和成本效益而言，ASIC在专用计算任务中表现最佳，其计算性能和能效远超通用GPU。但ASIC开发周期较长，且需达到一定生产规模才能实现成本优势。FPGA提供了一种介于GPU和ASIC之间的灵活解决方案，它的可编程性使硬件能够在算法迭代时进行有效优化，同时在开发周期上比ASIC更为短暂。

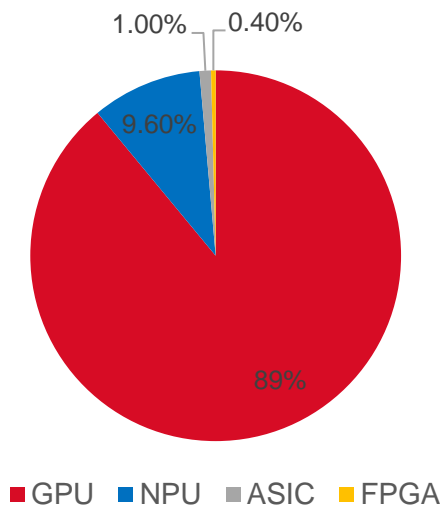
	GPU	FPGA	ASIC
定义	专为处理大量的并行任务而设计的处理器，通常用于图形渲染、数据并行计算以及深度学习等领域。	可在硬件层面重新配置以执行特定任务的集成电路，用于应对多变的计算需求。	专为执行一种特定应用或任务而设计和优化的集成电路，提供最高的性能和能效。
灵活性	高（软件层面的编程灵活性）	中到高（硬件可以重新配置，但需要硬件描述语言）	低（一旦设计完成，功能固定）
性能	高于FPGA在并行计算任务，但低于ASIC	依赖于特定应用，通常低于ASIC和GPU	高（针对特定应用优化）
功耗	高	低于GPU，但比ASIC高	低（高度优化）
成本	低到中（大规模生产）	中到高（需要时间进行配置和测试）	高（开发成本高，但大批量生产时单价低）
开发时间	短（软件开发）	中到长（需要硬件设计和测试）	长（设计、测试、优化周期长）
优点	并行处理能力强，软件生态丰富	灵活性较高，可以针对特定应用进行优化	针对特定任务有最高的性能和能效
劣势	功耗和成本较高，针对性不如ASIC和FPGA	开发周期和成本居中，性能不及ASIC	开发成本高，灵活性差，修改成本高
适用场景	图形渲染、深度学习、科学计算	快速原型设计、可变算法实现、信号处理	高性能计算任务、大规模生产的消费电子产品
代表企业/芯片（数据中心）	NVIDIA Tesla系列、AMD Instinct系列	Intel Stratix系列、Xilinx Virtex UltraScale系列	Google TPU（Tensor Processing Unit）、Amazon AWS Inferentia

资料来源：公开资料整理，东海证券研究所

2.1、AI芯片：GPU是AI芯片主流，AI芯片占AI服务器成本约70-75%

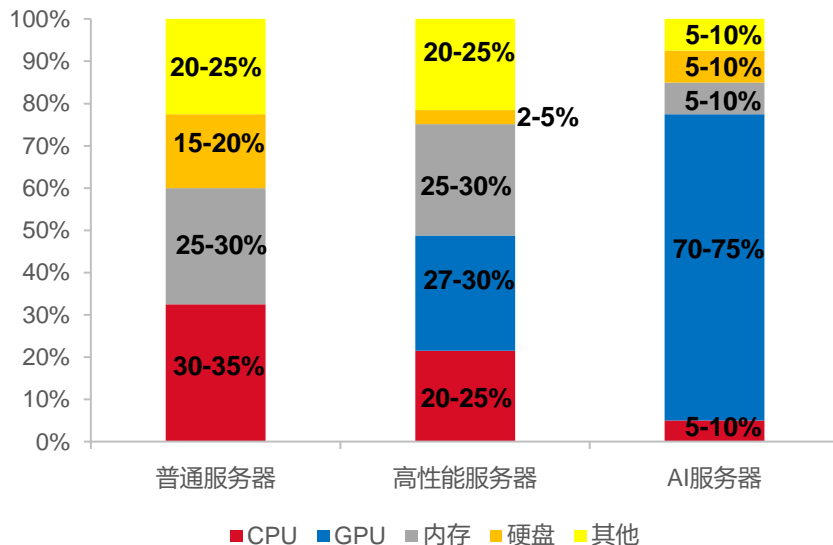
- **GPU是AI服务器的核心，约占近90%AI芯片市场份额，其价值量占AI服务器高达70-75%。**与传统服务器相比，AI服务器采用异构架构，能够搭载多个GPU、CPU及其他算力芯片来应对大规模并行计算的需求。传统服务器的CPU一般最多只有数十个核心，主要用来处理运算量较为复杂的数据。而GPU的具有数以千计的算术逻辑单元(ALU)和深度流水线，控制逻辑简单，省去了Cache的复杂性。因此在处理类型统一、相互无依赖的大规模数据时，GPU能够在一个无需中断的计算环境中高效运行。
- **GPU是机器学习的主流之选。**CPU由于受Cache和复杂的控制逻辑掣肘，导致在处理不同类型的数据时，需要引入分支和中断，增加了运算的复杂性和功耗。意味着在同等功耗下，GPU能效比显著高于CPU，能够加快AI模型训练和推理时间，从而减少机器学习模型从训练到部署的总时间。不仅如此，高性能GPU的制造工艺在英伟达和台积电等企业的领导下已趋向成熟，成本在AI芯片中具有优势，因此成为了市场主流之选。

2022中国AI芯片市场结构占比情况



资料来源：IDC，东海证券研究所

不同类型服务器成本构成示意图



资料来源：行行查研究中心，东海证券研究所

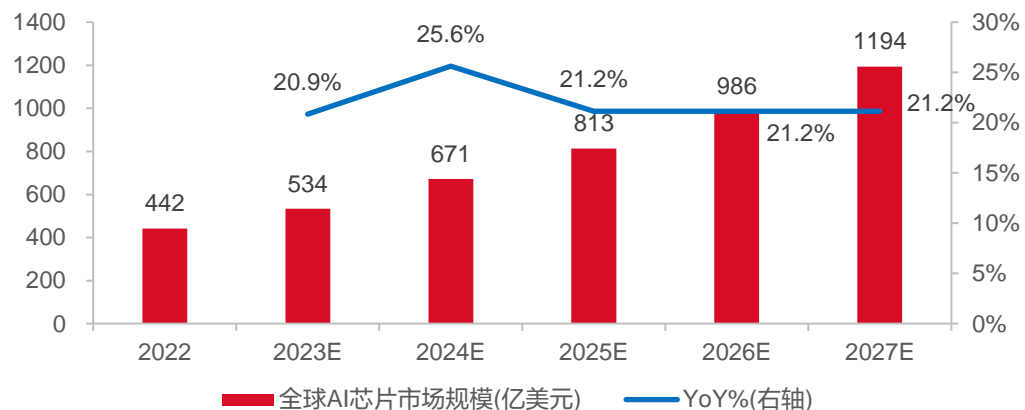
AI服务器与通用服务器的成本构成对比

零件类型	通用服务器成本构成		AI服务器成本构成	
	价格(美元)	BOM占比	价格(美元)	BOM占比
以2x Intel Sapphire Rapids Server为例				
CPU	1850	17.66%	5200	1.93%
8GPU + 4 NVSwitch Baseboard	-	-	195000	72.49%
内存DRAM	3930	37.52%	7860	2.92%
硬盘NAND	1536	14.66%	3456	1.28%
网卡SmartNIC	654	6.24%	10908	4.05%
机箱(外壳、背板、电缆)	395	3.77%	563	0.21%
主板	350	3.34%	875	0.33%
散热(散热器+风扇)	275	2.63%	463	0.17%
电源	300	2.86%	1200	0.45%
组装测试	495	4.73%	1485	0.55%
Markup	689	6.58%	42000	15.61%
总成本	10474	100.00%	269010	100.00%

资料来源：Semianalysis，东海证券研究所

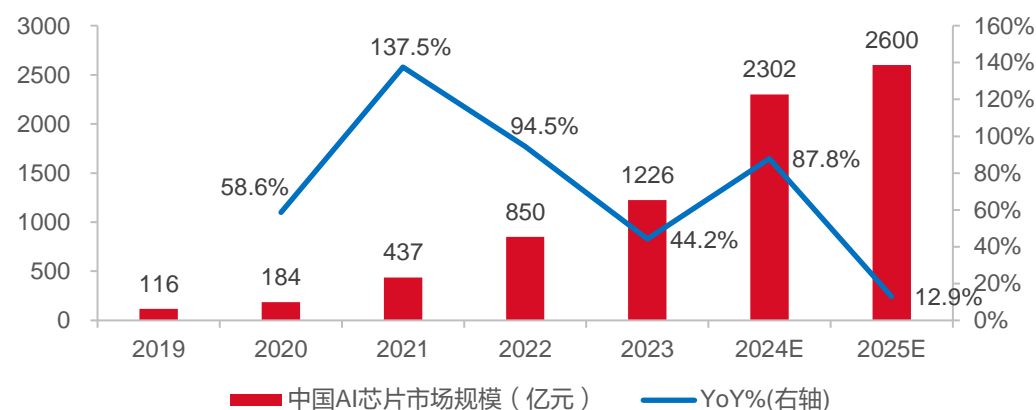
2.1、AI芯片：算力需求驱动AI芯片高增长，英伟达独霸鳌头

2022-2027E全球AI芯片市场规模及预测（亿美元）



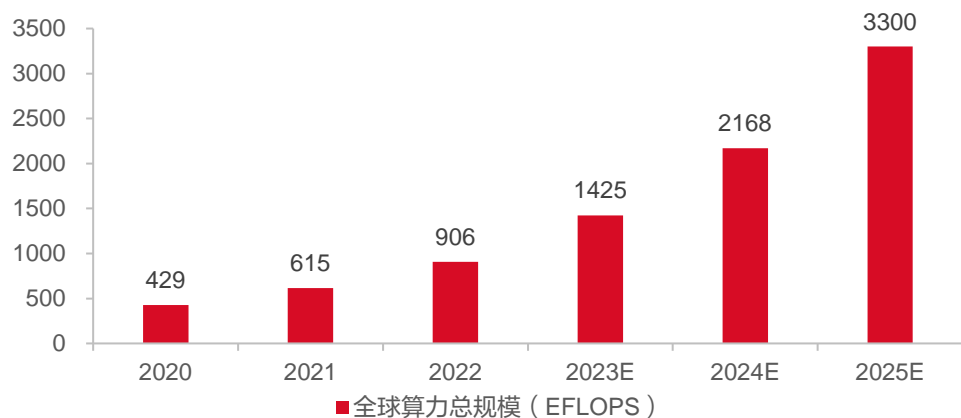
资料来源：Gartner，东海证券研究所

2019-2024E中国AI芯片市场规模及预测（亿元）



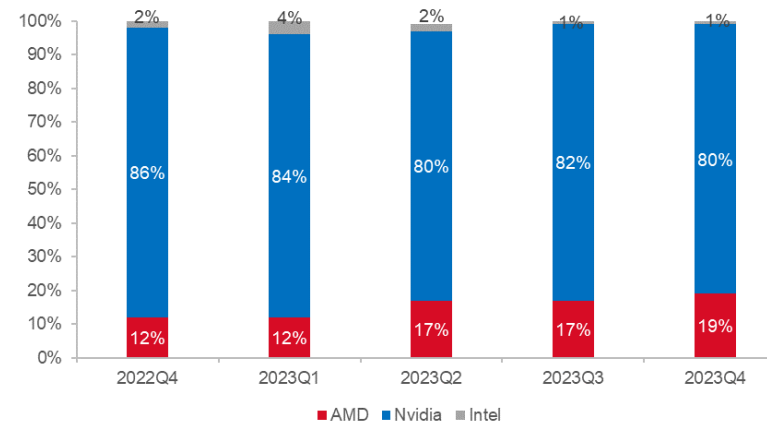
资料来源：中研普华产业院，东海证券研究所

2020-2025E全球算力总规模及预测(EFLOPS)



资料来源：信通院，IDC，东海证券研究所

2022Q4-2023Q4全球GPGPU芯片市场份额构成

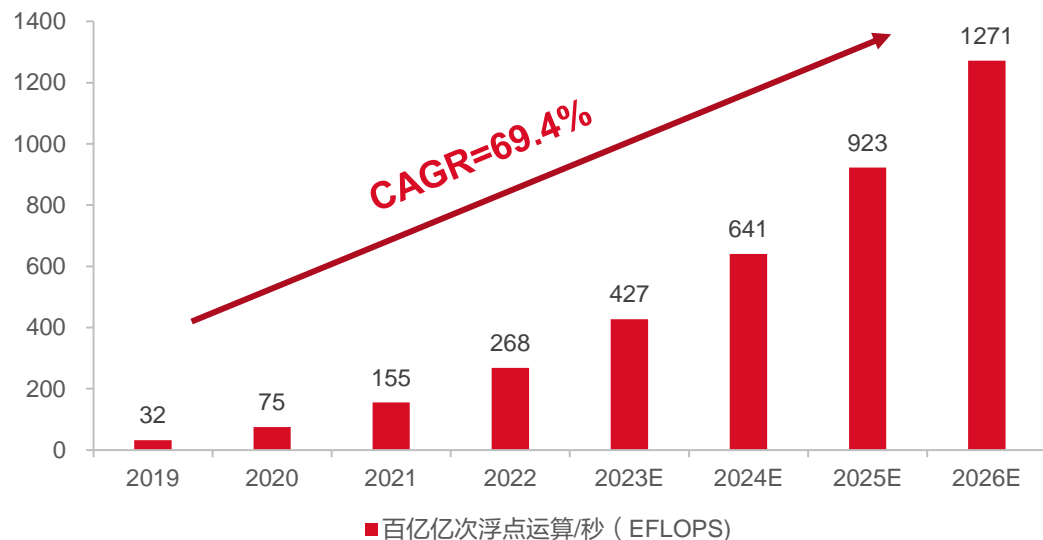


资料来源：Jon Peddie Research，东海证券研究所

2.2、AI芯片：我国AI芯片需求有望迎来爆发增长

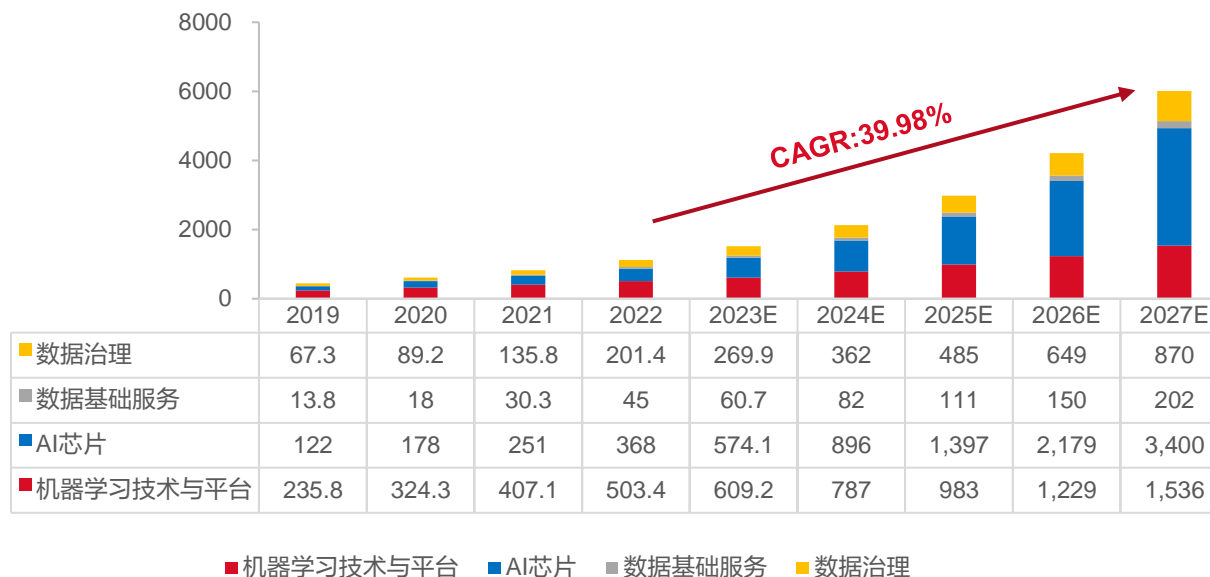
- AI大模型的迅速崛起带来了巨大的AI算力需求。AI大模型海量的数据处理、复杂的深度学习、多模态和跨领域的整合，以及广泛场景下实时性和交互性的要求，都对算力提出了更强的需求，这对我国的AI智能算力资源提出了挑战。
- 随着国产AI大模型如雨后春笋般涌现，智能算力缺口与日俱增，AI芯片需求有望迎来爆发增长。2023年以来，文心一言、讯飞星火、通义千问等上百家国产大模型争相涌现，参数规模从几亿乃至上万亿，广泛应用于云计算、数据中心、边缘计算、消费电子、智能制造、智能驾驶、智能金融及智能教育等领域，用于AI训练和推理的智能算力缺口与日俱增，AI芯片需求持续旺盛。根据Frost&Sullivan，2022年我国AI芯片市场规模达到368亿元，预计到2027年，市场规模将进一步扩大至3400亿元，2022-2027E CAGR将有望达到39.98%。

2020-2025E中国智能算力规模及预测（EFLOPS）



资料来源：IDC，浪潮信息，东海证券研究所

2019-2027E中国AI基础设施市场规模及预测（亿元）



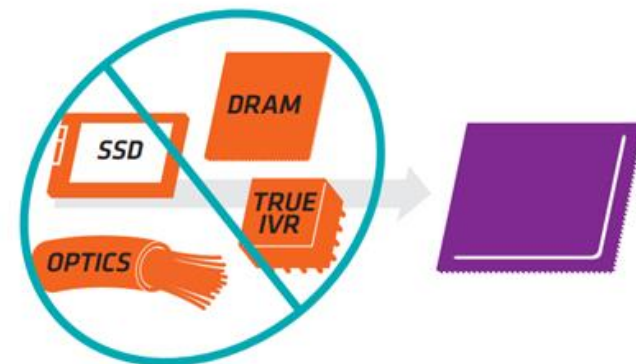
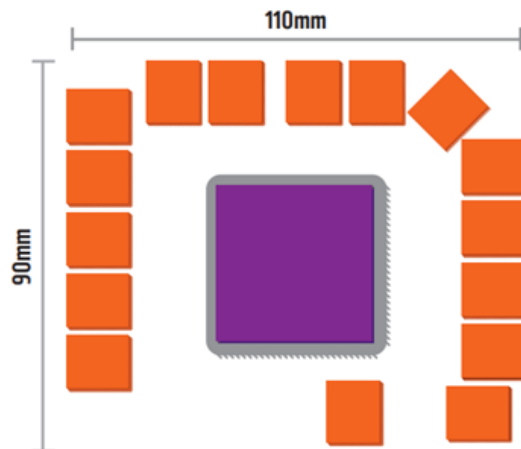
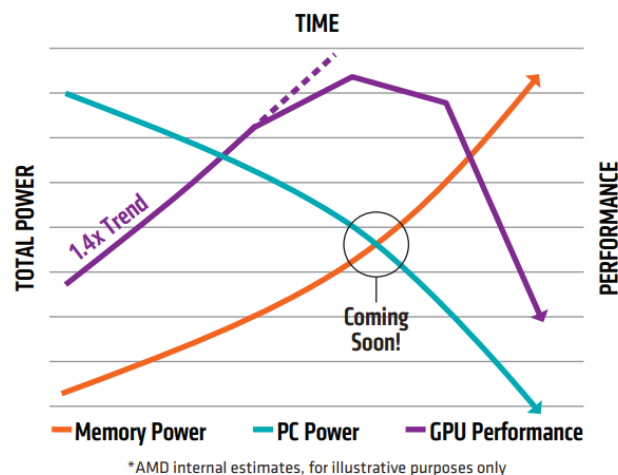
资料来源：弗若斯特沙利文，东海证券研究所

2.2、HBM存储模组：解决了AI发展的存储墙问题

问题1：GDDR速率赶不上CPU发展速率

问题2：显卡面积限制GDDR数量

问题3：高度集成化技术难度增大

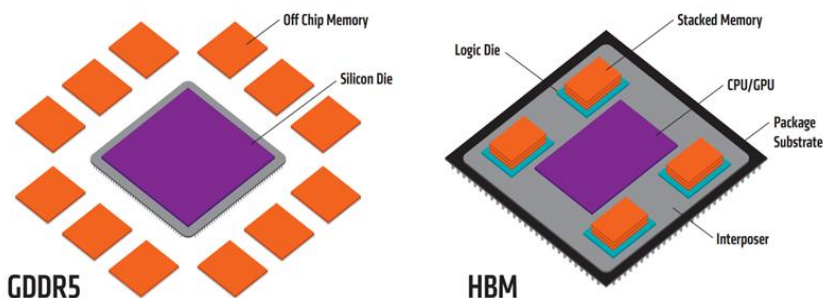


资料来源：CSDN，AMD，东海证券研究所

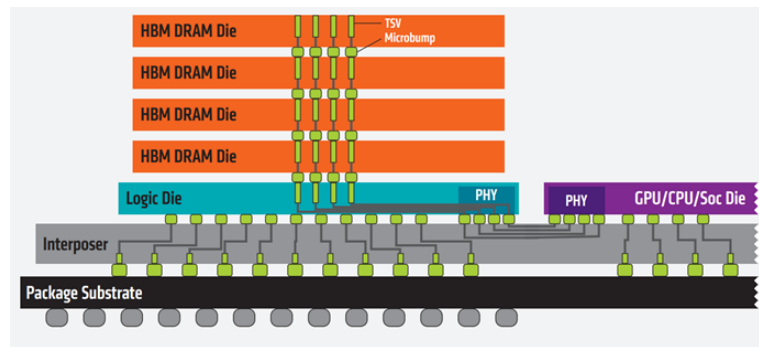
- **GDDR是显示用内存**：Graphics Double Data Rate的缩写，为显存的一种，GDDR是为了设计高端显卡而特别设计的高性能DDR存储器规格，其有专属的工作频率、时钟频率、电压，因此与市面上标准的DDR存储器有所差异，与普通DDR内存不同且不能共用，GDDR产品具备高带宽、低延时、低功耗、高稳定性等特征。GDDR广泛应用于显卡、游戏主机和其他需要高性能图形处理的设备上。
- **内存墙问题1**：近几十年来，处理器的性能以每年大约55%速度快速提升，而内存性能的提升速度则只有每年10%左右。不均衡的发展速度造成了当前内存的存取速度严重滞后于处理器的计算速度，内存瓶颈导致高性能处理器难以发挥出应有的功效。
- **内存墙问题2**：随着GPU性能不断提升，匹配GPU的GDDR数量越来越多，而GPU的面积规格有限，导致显卡的体积越来越大，后期的散热与产品规格问题日益严重。
- **内存墙问题3**：随着对计算性能要求越来越高，各个分离器件芯片的集成趋势要求越来越高，但是难度越来越大。最后，先进封装技术不断迭代发展，与GPU高度融合的HBM产品应运而生。

2.2、HBM存储模组：带宽、位数、体积、功耗显著优于GDDR5

GDDR5与HBM的产品框架图

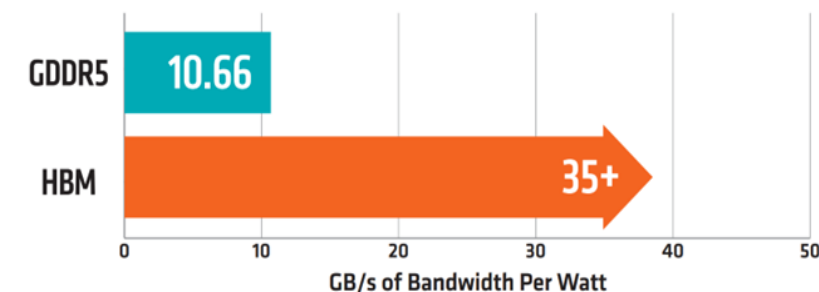


HBM的垂直截面图



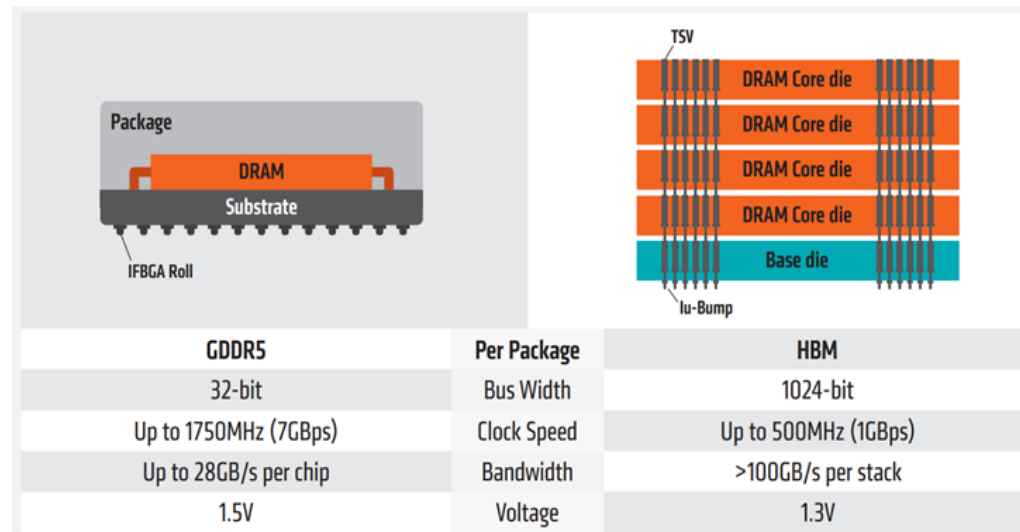
HBM相对GDDR的性能优势显著

HBM vs GDDR5:
Better bandwidth per watt¹

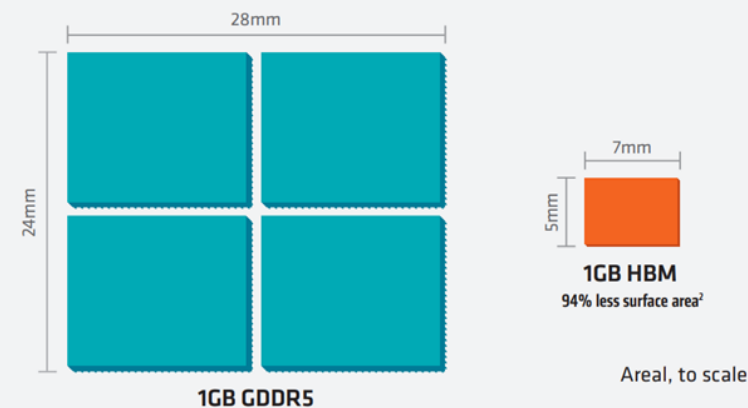


- **HBM 技术简介：** HighBandwidth Memory，即高带宽内存，是一种新兴的DRAM 解决方案。HBM具有基于TSV(硅通孔)和芯片堆叠技术的堆叠 DRAM 架构，通过 uBump 和 Interposer(中介层，起互联功能的硅片)实现超快速连接。Interposer 再通过 Bump 和 Substrate(封装基板) 连接到 BALL，最后 BGA BALL 再连接到 PCB 上。
- **HBM 优势：** (1) 极高带宽：达到 1T/s；(2) 体积减小：比 GDDR 降低 94% 的尺寸；(3) 低功耗：高度集成后拥有比 GDDR 更小的电压与功耗。

HBM与GDDR的性能参数对比



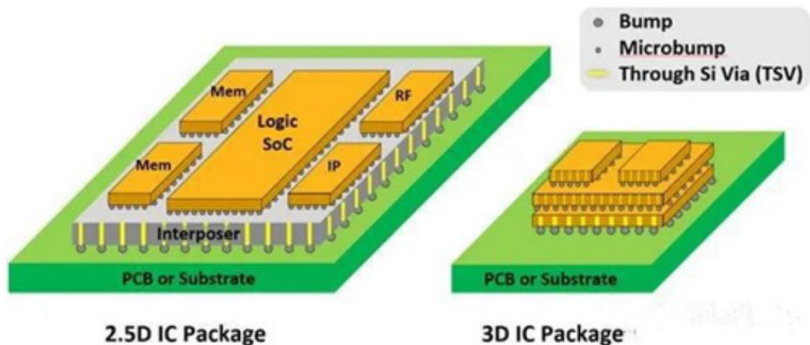
HBM vs GDDR5:
Massive space savings



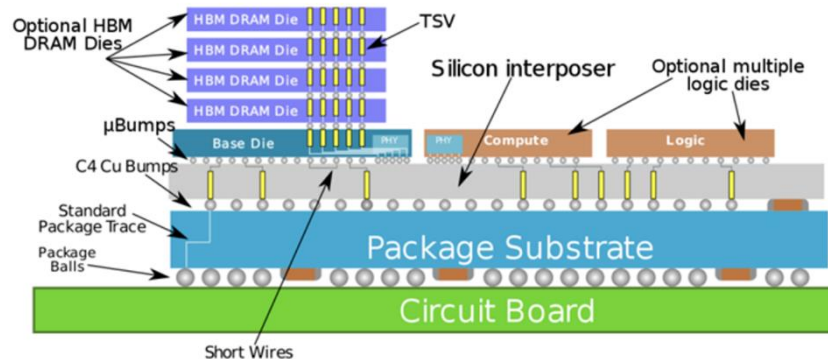
资料来源：CSDN，AMD，东海证券研究所

2.2、HBM存储模组：需要2.5D/3D、TSV、uBUMP等先进封装技术

2.5D与3D封装及关键技术图



台积电用2.5D的COWOS技术将HBM与逻辑芯片连接图

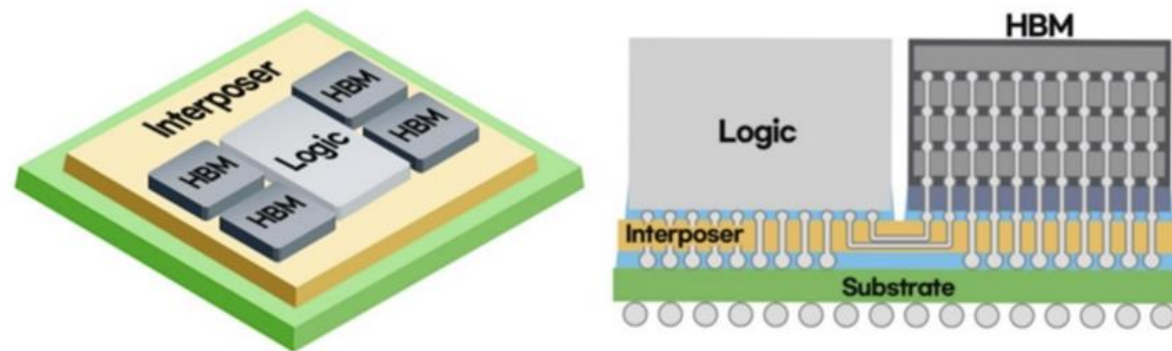


资料来源：电子技术应用，ANSYS，东海证券研究所

资料来源：台积电，与非网，东海证券研究所

TSV为HBM实现的关键技术之一

三星用2.5D的I-Cube技术将HBM与逻辑芯片连接图



资料来源：SK海力士，东海证券研究所

资料来源：三星，与非网，东海证券研究所

2.2、HBM存储模组—2024年三大厂商量产带宽高达1TB/s以上HBM3E

HBM迭代产品参数演变

HBM迭代产品参数演变					
产品名称	芯片密度	带宽	堆叠高度	容量	I/O速率
HBM1	2Gb(4-hi)	128GB/s	4层	1GB	1Gbps
HBM2	8Gb	307GB/s	4/8层	4/8/16GB	2.4Gbps
HBM2E	8Gb/16Gb	460GB/s	4/8层	8/16GB	3.6Gbps
HBM3	16Gb	819GB/s	8/12层	16/24GB	6.4Gbps
HBM3E	NA	1.2TB/s	8层	24GB	9.2Gbps

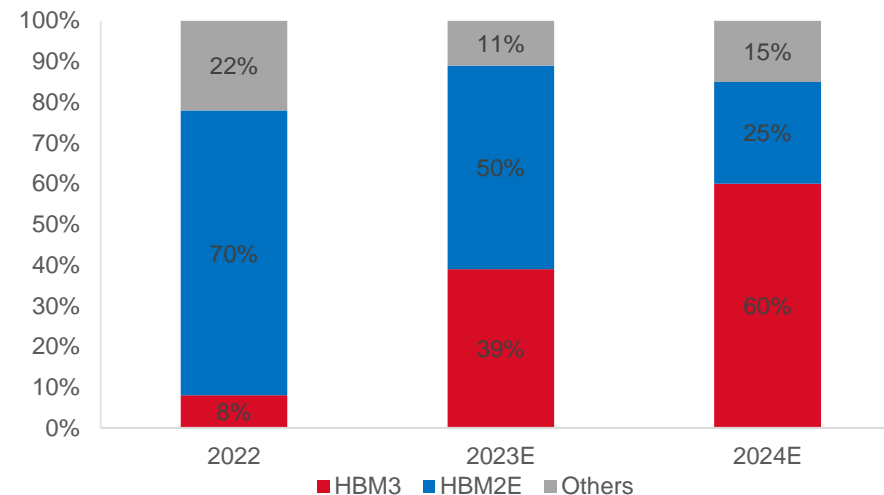
资料来源：美光科技，SK海力士，三星，东海证券研究所

全球大厂HBM技术路线图

企业名称	发布时间	产品	量产时间
SK海力士	2014年	HBM1 (SK海力士、AMD)	NA
	2018年	HBM2	NA
	2019年8月	HBM2E	2020年7月
	2021年10月	HBM3	2022年6月
	2023年8月	HBM3E	2024年上半年
三星	2016年6月	HBM2	2018年
	2020年2月	HBM2E	2020年
	2021年2月	HBM-PIM (存算一体)	2021年完成验证
	-	HBM3	2022年
	2024年2月	HBM3E (36GB)	2024年上半年
美光科技	2023年	HBM3E (24GB)	2024年2月26日

资料来源：电子产品世界，SK海力士，三星，东海证券研究所

2022-2024年HBM2E与HBM3比重转进预估



资料来源：同花顺财经，TrendForce，东海证券研究所

- **HBM技术更新快速**：2024年三大国际大厂都将主要量产HBM3与HBM3E，HBM3产品占有率将达到60%以上。
- **HBM产品壁垒或将导致高端HBM被国际大厂垄断**：（1）2.5D或3D封装的技术与产能，台积电的COWOS更新到第5代，产能供不应求；（2）高端GPU芯片的供货，HBM与GPU封装在一起组成AI加速卡；（3）GDDR原厂颗粒供货，一般19nm及以下制程颗粒，甚至到1α，1β制程的高制程存储颗粒芯片。

2.2、HBM存储模组：全球HBM需求规模预测高速增长

全球HBM历年总规模测算

年份	2022	2023E	2024E	2025E	2026E
全球AI服务器出货量预测（万台）	85.5	118.2	150.4	189.5	236.9
平均单台服务器搭载的GPU个数	4	4	6	8	8
单个GPU的HBM存储容量（GB）	40	60	80	90	100
HBM单价及预测：美元/GB	18	16	15	12	11
全球HBM规模测算：亿美元	24.62	45.39	108.29	163.73	208.47

资料来源：TrendForce，Gartner，浪潮，东海证券研究所

2022-2024全球HBM供给格局及估算

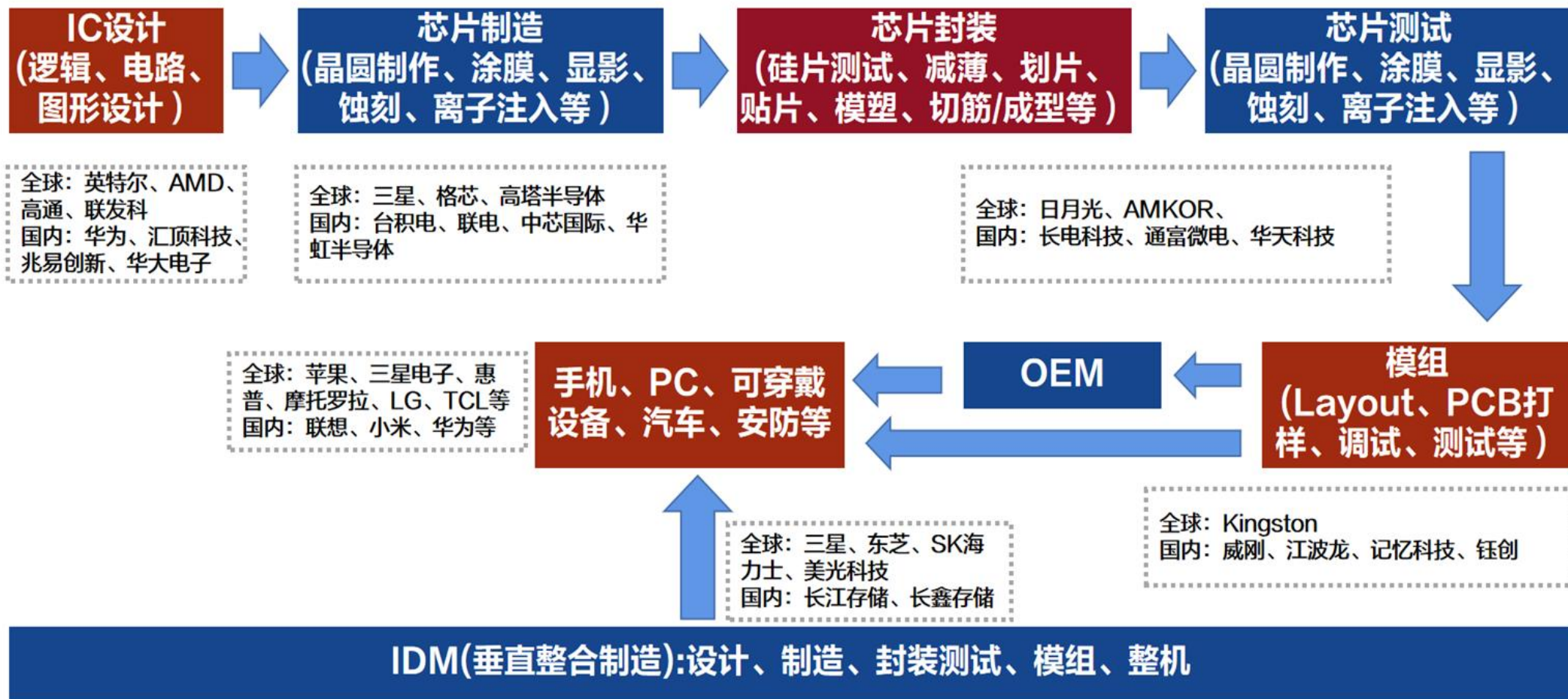
公司	2022	2023E	2024E
SK海力士	50%	46-49%	47-49%
三星	40%	46-49%	47-49%
美光科技	10%	4-6%	3-5%

资料来源：TrendForce，东海证券研究所

- HBM历年全球需求：**根据TrendForce数据，全球AI服务器2022年约为85.5万台，到2026年约为236.9万台，CAGR为29%。平均单个服务器搭载的GPU有2/4/8/16个，2024年预计多为4/8个。单个GPU搭载HBM个数若干，总容量在60-100GB之间，随着服务器性能不断升级，容量会逐步增大。市场HBM产品价格理论上受到技术不断进步，产能扩张影响，平均价格小幅度下滑。我们预计2024年HBM全球需求或将达到108.29亿美元。
- HBM全球历年供给：**根据TrendForce数据，未来3年全球的HBM主要供应链依然是SK海力士、三星、美光占据。由于先进制程GPU、GDDR颗粒、2.5D/3D封装技术与产能局限，我们认为全球高端HBM呈现寡头垄断格局。
- 国产HBM需求与供给：**随着全球AI的发力发展，国产AI服务器不断追赶，对HBM的需求依然存在。目前存储芯片厂如长鑫、长存，封测厂如通富微电、长电科技等均在逐步布局中。

2.2、HBM存储模组：国内存储企业逐步布局HBM

全球存储产业核心环节产业链结构图

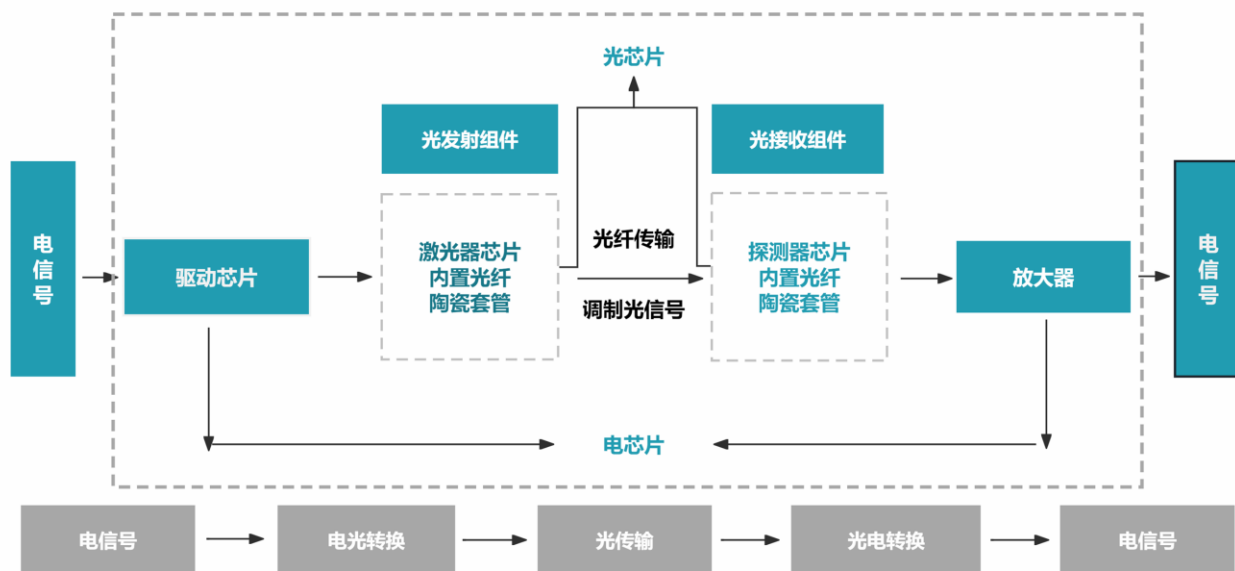


资料来源：兆易创新，同花顺，东海证券研究所

2.3、光模块：服务器互连和数据中心互连的核心器件

- 光模块(Optical Module)是进行光电和电光转换的设备。由光电子器件（光发射组件和光接收组件）、功能电路和光接口等组成。光模块在发送端把电信号转换成光信号，通过光纤传送后，接收端再将光信号转换成电信号。
- 光模块可按照功能、传输速率、复用技术、适用光纤类型和封装形式等标准分类。按照传输速率分类，目前主要有100G、200G、400G、800G、1.6T等；按照功能分类，光模块可分为光接收模块，光发送模块，光收发一体模块和光转发模块，一般特指光收发一体模块；按照封装形式分类，常见的有SFP，SFP+，SFF，千兆以太网路界面转换器（GBIC）等。

光模块基本工作原理



资料来源：联特科技招股书，东海证券研究所

光模块基本分类和对应特征

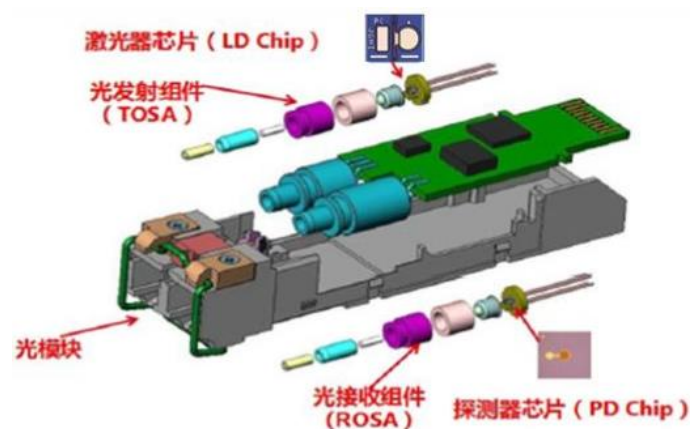
分类标准	光模块类别	特征
传输速率	155Mb/s、622Mb/s、1.25Gb/s、2.5Gb/s、2.97Gb/s、4.25Gb/s、6.1Gb/s、8.5Gb/s、10Gb/s、25Gb/s、40Gb/s、100Gb/s、200Gb/s、400Gb/s、800Gb/s、1.6Tb/s等	指每秒传输比特数，通常传输速率越高，代表的技术难度越高；光模块的发展方向之一是高传输速率
复用技术	时分复用系统 850nm、1310nm、1550nm等波段	850nm波段用于多模光纤传输，传输距离短，多用于2km以内短距离传输 1310nm波段用于单模光纤传输，传输损耗大色散小，一般用于40km以内的传输 1550nm波段用于单模光纤传输，传输损耗小色散大，一般用于40km以上的长距离传输，最远可以无中继直接传输120km
复用技术	WDM（波分复用）系统 CWDM系列（粗波分复用） DWDM系列（密集波分复用）	使用20nm间隔的波长，将多个波长的光信号复用进一根光纤内传送数据 使用0.4nm或者0.8nm间隔的波长，将多个波长的光信号复用进一根光纤内传送数据
适用光纤类型	单模光纤 多模光纤	纤芯较细，只能传输一种模式的光，适用于远程通讯 纤芯较粗，可传输多种模式的光。多模光纤模间色散较大，适用于短距离通讯
封装形式	SFP、SFP+、XFP、SFP28、QSFP+、QSFP28、QSFP-DD、OSFP等	光模块的封装形式呈多样化，满足行业标准组织的多源协议（MSA）

资料来源：联特科技招股书，东海证券研究所

2.3、光模块：TOSA占据光模块成本结构的35%，高端光芯片国产化率较低

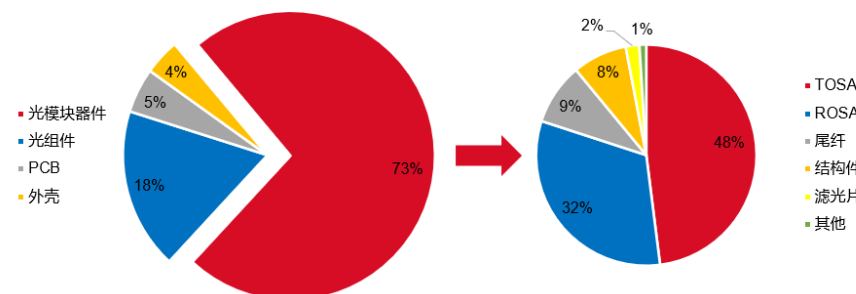
- 从成本结构看，光模块器件占据了光模块73%的成本，此外按照成本大小依次为光组件、PCB（印刷电路板）和外壳。光模块器件成本结构中，以激光器为主的光发射组件（TOSA）和以探测器为主的光接收组件（ROSA）分别占据了48%和32%的成本，光发射组件（TOSA）占据光模块成本的35%。
- 高端光芯片（25G以上）国产替代率较低。国内企业在2.5G和10G光芯片领域基本实现了核心技术的掌握，国产化率分别为90%和60%，但是25G光芯片国产化率为20%，25G以上光芯片国产化率仅为5%，国产替代空间较大。

光模块结构示意图（SFP+封装）



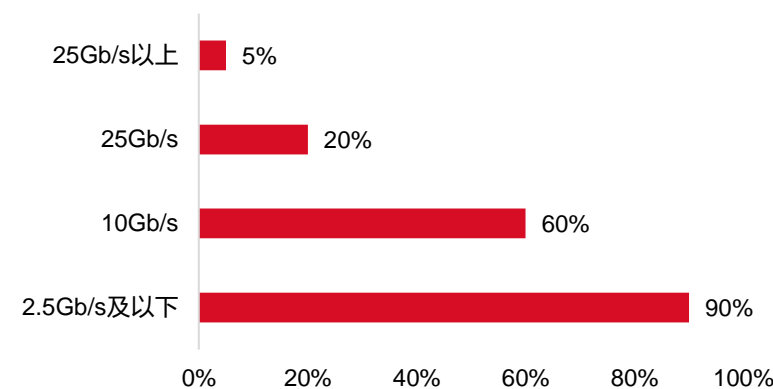
资料来源：源杰科技招股书，东海证券研究所

光模块和光器件成本结构



资料来源：华经情报网，东海证券研究所

25G及以上光芯片国产化率较低

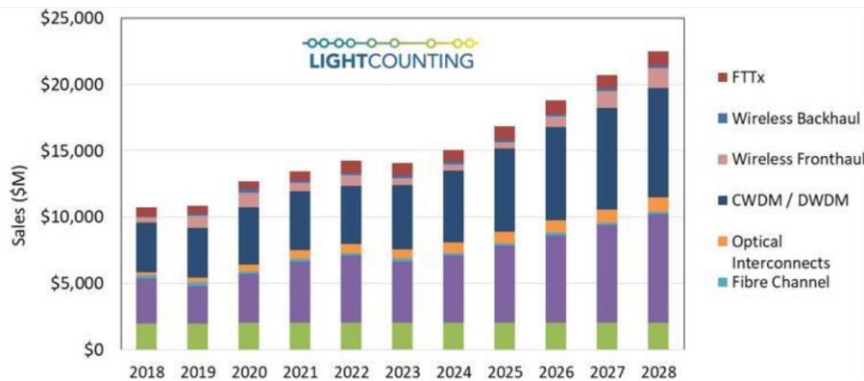


资料来源：ICC、中商情报网，东海证券研究所

2.3、光模块：AI驱动光模块市场高速扩张，高传输速率光模块成主流

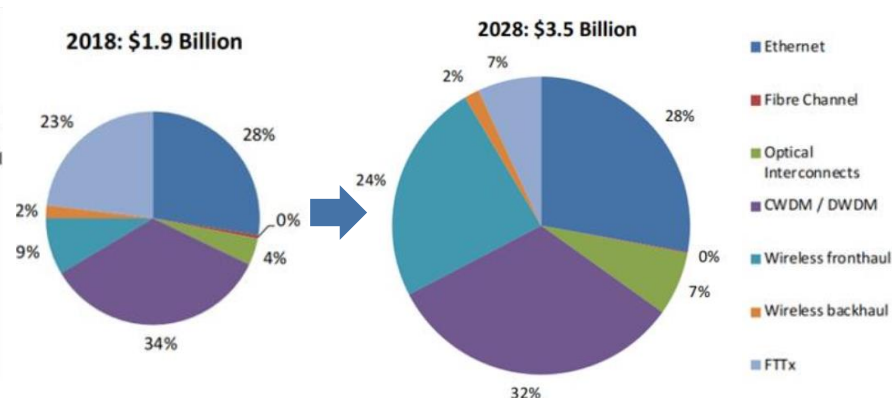
- **全球视角：**以ChatGPT为代表的生成式AI工具正引领新一轮科技革命，AI军备竞赛的开启大幅拉动了算力的爆发式需求。前沿科技产业化的落地需要云厂商庞大的算力支持，而光通信网络是算力网络的重要基础和坚实底座，这将进一步推动海外云巨头对于数据中心硬件设备的需求增长与技术升级。Lightcounting预测，全球光模块的市场规模在未来5年将以11%的年复合增长率持续上升，2027年将突破200亿美元。2023年开始，800G有望拉动新一轮增长，预计在2026年突破30亿美元大关。
- **国内视角：**2023年是经济全面复苏的重要一年，数字经济成为推动经济增长的重要引擎，数字中国顶层设计的落地将带来算力提升能耗增长。同时伴随着“东数西算”战略的逐步落地，国内数据中心也同步加快新建、扩容步伐，而光模块作为数据中心内部设备互联的载体，在加大AI投入的背景下，长期来看光模块市场有望持续扩张。国家统计局数据显示，2023年上半年，新型基础设施建设投资同比增长16.2%，其中5G、数据中心等信息类新型基础设施投资增长13.10%，工业互联网、智慧交通等融合类新型基础设施投资增长34.10%。Lightcounting预计2028年中国光模块市场规模有望达35亿美元，其中光波分复用技术（DWDM）占比32%，以太网占比28%。
- **AI算力需求的高速扩张对数据中心的吞吐量有更高的要求。**从传输速率角度看，800G甚至1.6T光模块成为未来光模块市场主流需求。

全球光模块市场未来五年CAGR达11%



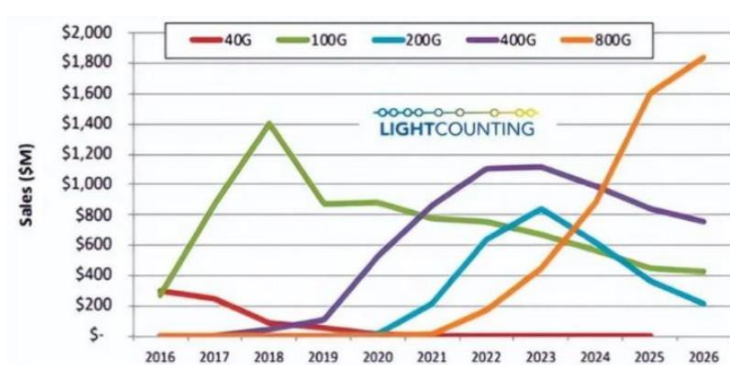
资料来源：Lightcounting，中际旭创公告，东海证券研究所

2028年中国光模块市场规模有望达35亿美元



资料来源：Lightcounting，中际旭创公告，东海证券研究所

800G光模块市场前景广阔

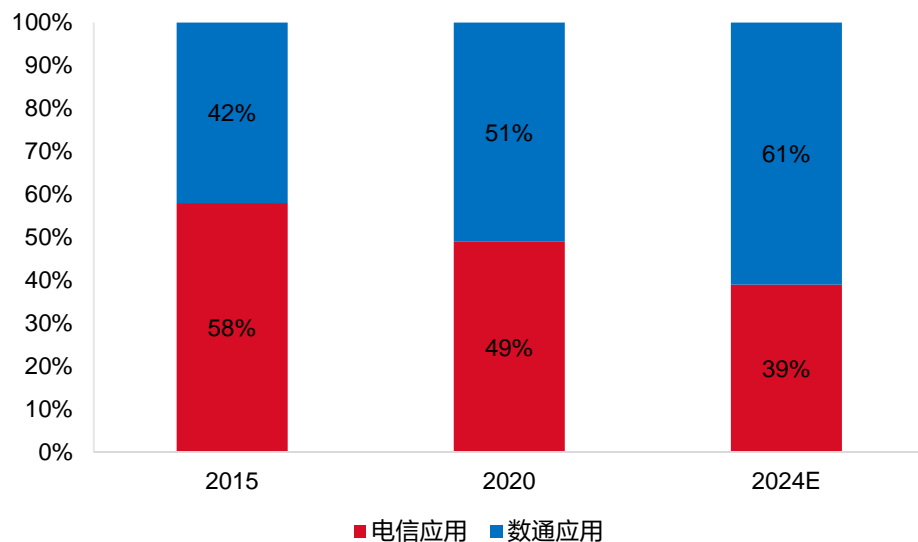


资料来源：Lightcounting，未来智库，东海证券研究所

2.3、光模块：数通应用占比迅速上升，国内企业占据半壁江山

- 按照光模块下游应用分类，数通应用将成为未来主流应用领域。2015年，光模块下游领域中，数通市场占比42%，电信应用占比58%。2020年，数通市场占比上升至51%。受益于AI浪潮下的数据中心建设，数通市场是光模块下游应用领域中增速最快的市场，主要包括云计算、大数据等。据预测，2024年全球光模块在数通市场、电信市场的应用占比分别为61%、39%。
- 中国企业在光模块市场中逐步掌握话语权，2022年全球前十大光模块厂商国内企业已占据半壁江山。根据Lightcounting，2016年全球前十大光模块厂商国内企业只有三家，市场基本由美国企业主导。到2022年，国内企业已在前十大厂商中占据一半的席位，其中中际旭创和Coherent并列第一，其他四家分别为华为、光迅科技、海信和新易盛。

全球光模块下游应用占比



资料来源：中商情报网，东海证券研究所

全球前十大光模块厂商

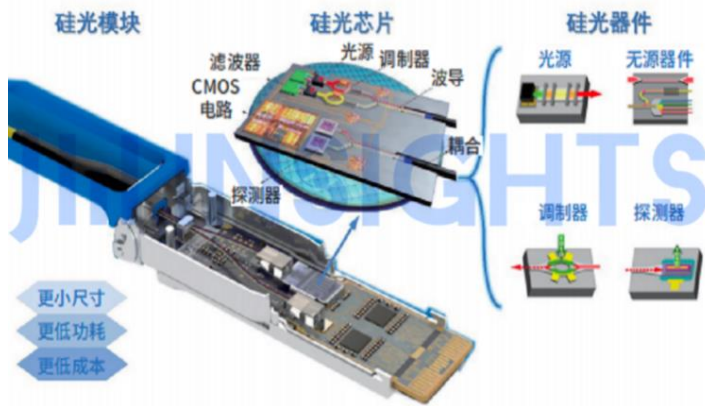
排名	2016	2018	2021	2022
1	Finisar (美)	Finisar (美)	II-VI (美) & 中际旭创 (中) 并列	中际旭创 (中) & Coherent (美) 并列
2	海信 (中)	中际旭创 (中)	中际旭创 (中) 并列	中际旭创 (中) 并列
3	光迅科技 (中)	海信 (中)	华为 (海信) (中)	Cisco (Acacia) (美)
4	Acacia (美)	光迅科技 (中)	Cisco (Acacia) (美)	华为 (海信) (中)
5	FOIT (Avago) (美)	FOIT (Avago) (美)	海信 (中)	光迅科技 (中)
6	Oclaro (美)	Lumentum/Oclaro (美)	光迅科技 (中)	海信 (中)
7	中际旭创 (中)	Acacia (美)	Broadcom (美)	新易盛 (中)
8	Sumitomo (日)	Intel (美)	HGG (美)	HGG (美)
9	Lumentum (美)	Aoi (美)	新易盛 (中)	Intel (美)
10	Source Photonics (美)	Sumitomo (日)	Molex (美)	Source Photonics (美)

资料来源：Lightcounting，中际旭创公告，东海证券研究所

2.3、光模块：高集成度、低成本、低能耗的硅光技术是未来发展趋势之一

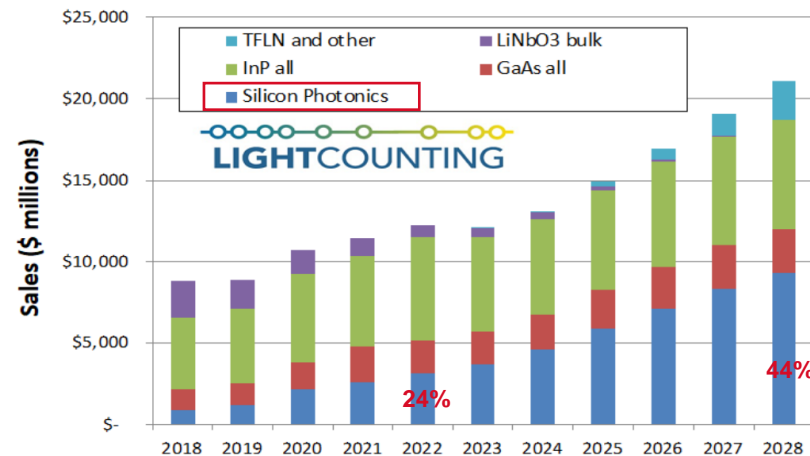
- 硅光技术是光模块未来的重要发展方向之一。硅光解决方案集成度高，成本低，传输带宽高，同时在峰值速度、能耗等方面均具有良好表现。硅光子技术是基于硅和硅基衬底材料（SiGe/Si、SOI等），利用现有CMOS工艺进行光器件开发和集成的新一代技术。鉴于良率和损耗问题，硅光模块方案的整体优势尚不明显，但在超400G的短距场景、相干光场景中，传统DML和EML成本较高，硅光模块的低成本优势或使得其成为数据中心网络向400G升级的主流产品。
- 数据中心是硅光子技术的主要应用领域。根据Lightcounting，光通信行业已经处在硅光技术SiP规模应用的转折点，使用基于SiP的光模块市场份额将从2022年的24%增加到2028年的44%。

硅光模块结构图



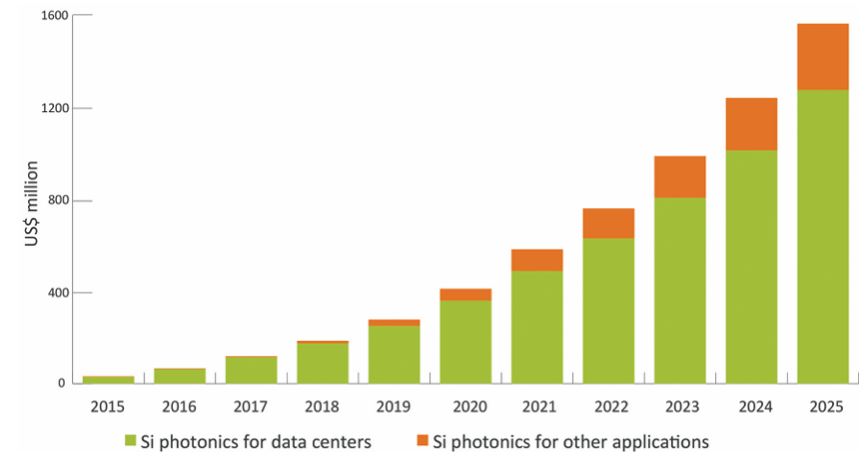
资料来源：集微咨询，东海证券研究所

全球硅光模块市场规模逐年上升



资料来源：Lightcounting，中际旭创公告，东海证券研究所

硅光子技术主要应用于数据中心

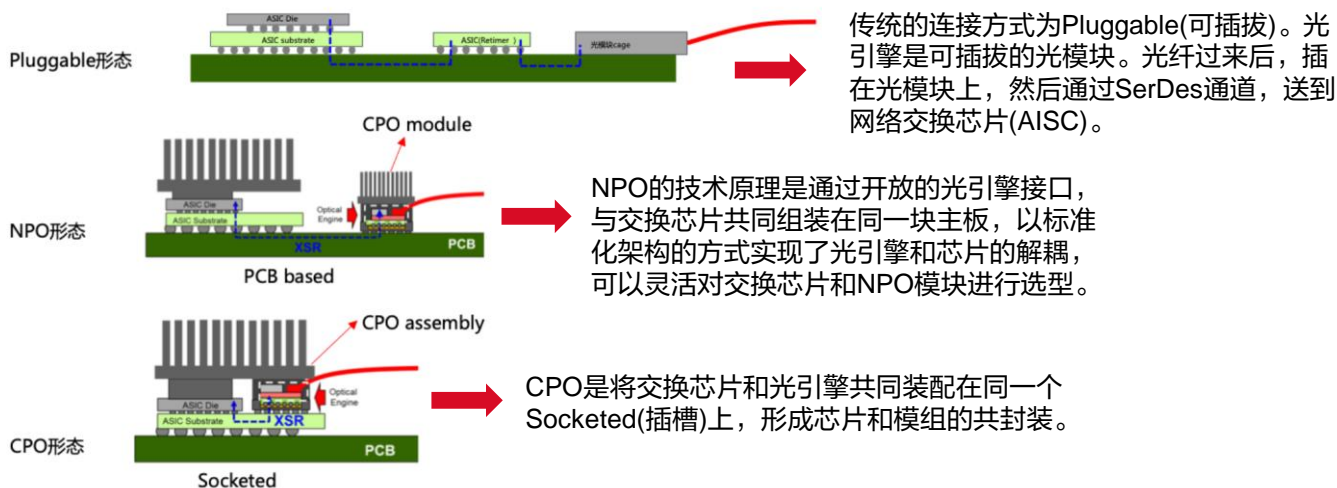


资料来源：《硅光子技术及产业发展研究》，吴冰冰，东海证券研究所

2.3、光模块：CPO方案有望在高速光模块中广泛应用

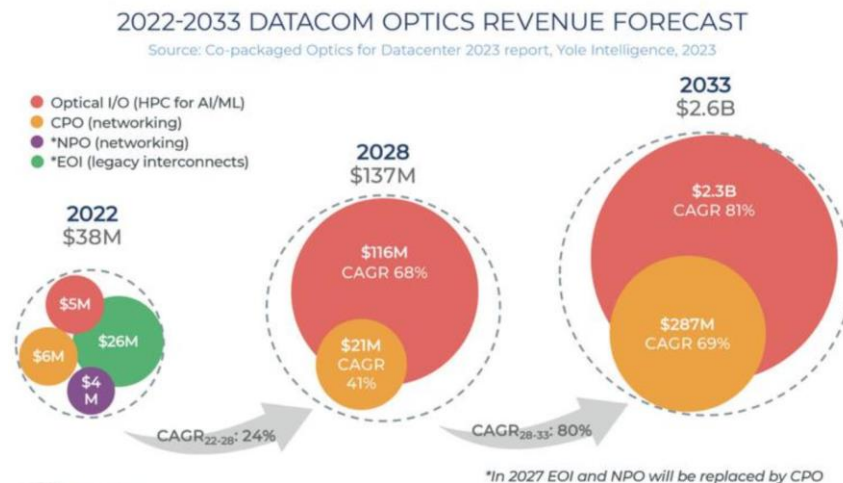
- **CPO方案将主要应用于800G及以上的光模块中。**光电共封装（CPO）指的是交换ASIC芯片和硅光引擎在同一高速主板上协同封装，从而降低信号衰减、降低系统功耗、降低成本和实现高度集成。CPO处于发展的起步状态，其行业标准形成预计还要一定时间，但CPO的成熟应用或会带来光模块产业链生态的重大变化。硅光技术既可以用在传统可插拔光模块中，也可以用在CPO方案中。800G传输速率下硅光封装渗透率会有提升，而CPO方案则更多的是技术探索。但从1.6T开始，传统可插拔速率升级或达到极限，后续光互联升级可能转向CPO和相干方案。LightCounting表示，AI对网络速率的需求是目前的10倍以上，在这一背景下，CPO有望将现有可插拔光模块架构的功耗降低50%，有效解决高速高密度互联传输场景，并预计CPO出货预计将从800G和1.6T端口开始，于2024至2025年开始商用，2026至2027年开始规模上量，主要应用于超大型云服务商的数通短距场景。Yole数据显示，2022年CPO市场产生的收入约为3800万美元，预计2033年将达到26亿美元，2022-2033年复合年增长率为46%。
- **LPO（线性驱动可插拔）技术强调“可插拔”，区别于CPO方案中光模块的不可插拔。**LPO与传统光模块的主要区别在于线性驱动，其将光模块中的DSP/CDR芯片取出，将相关功能集成到设备侧的交换芯片中，具有低功耗、低成本、低延迟和易于维护的特点。

CPO图解



资料来源：锐捷官网，东海证券研究所

CPO市场未来将高速扩张



资料来源：Yole，中际旭创公告，东海证券研究所

2.3、光模块：上游零部件供应商较为分散，下游主要为运营商和设备商

光模块位于光通信产业链的中游。主要厂商包括中际旭创、上游主要为光芯片、光器件、PCB、结构件、外壳、电芯片等零部件，下游主要面向电信运营商、数据中心运营商、通讯设备厂商等。光模块行业上游原材料供应充足，产业发展成熟，供应商议价能力适中。



资料来源：工信部、中商情报网、前瞻经济学人，东海证券研究所

目录

第一部分 大模型带动AI服务器高增长

第二部分 算力芯片与光模块长期受益

第三部分 A股上市公司代表

第四部分 投资建议

第五部分 风险提示

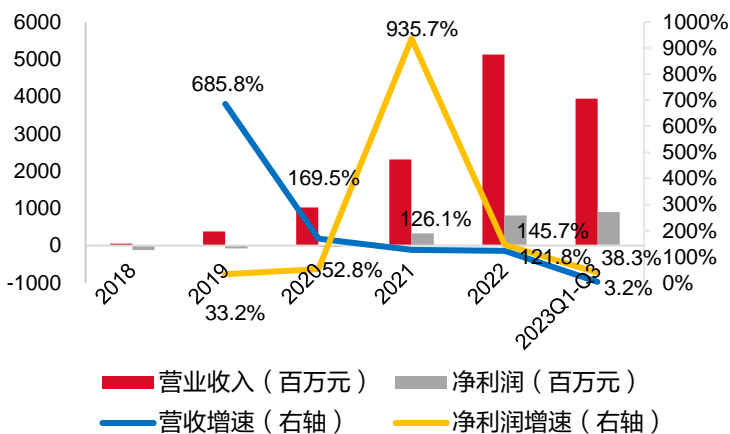


3.1、海光信息：深耕CPU、DCU领域，国产龙头持续受益于AI产业浪潮

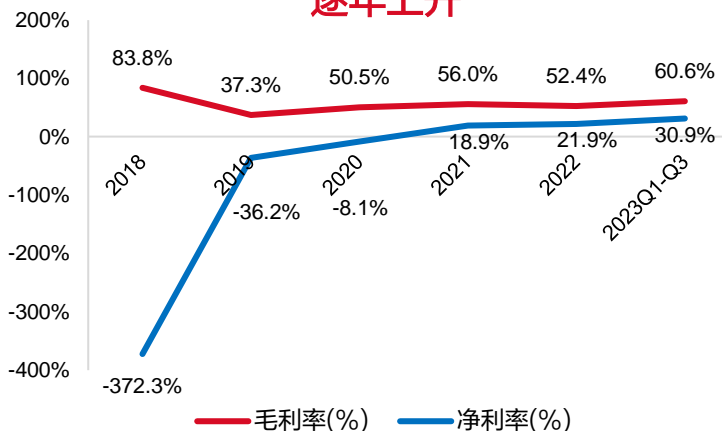
- 海光信息成立于2014年，2022年在科创板上市(688041.SH)，主要从事研发、设计和销售应用于服务器、工作站等计算、存储设备中的高端处理器。公司秉承“销售一代、验证一代、研发一代”的产品研发策略，产品主要包括海光通用处理器（CPU）和海光协处理器（DCU）。海光CPU兼容市场主流的x86指令集，是当前生态兼容性最优异的芯片之一，完全满足商业市场需求。
- 海光DCU深算系列属于GPGPU的一种，采用“类CUDA”通用并行计算架构，能够较好适配、适应国际主流商业计算软件和AI软件，产品性能达到国内领先。目前主要部署在服务器集群或数据中心，为应用程序提供性能高、能效比高的算力，支撑高复杂度和高吞吐量的数据处理任务。

产品类型	处理器种类	指令集	主要产品	产品特征	典型应用场景
海光CPU	通用处理器	兼容x86指令集	海光3000系列 海光5000系列 海光7000系列	内置多个处理器核心，集成通用的高性能外设接口，拥有完善的软硬件生态环境和完备的系统安全机制，适用于数据计算和事务处理等通用型应用	云计算、物联网、信息服务等
海光DCU	协处理器	兼容“类CUDA”环境	海光8000系列	内置大量运算核心，具有较强的并行计算能力和较高的能效比，适用于向量计算和矩阵计算等计算密集型应用	大数据处理、人工智能、商业计算等

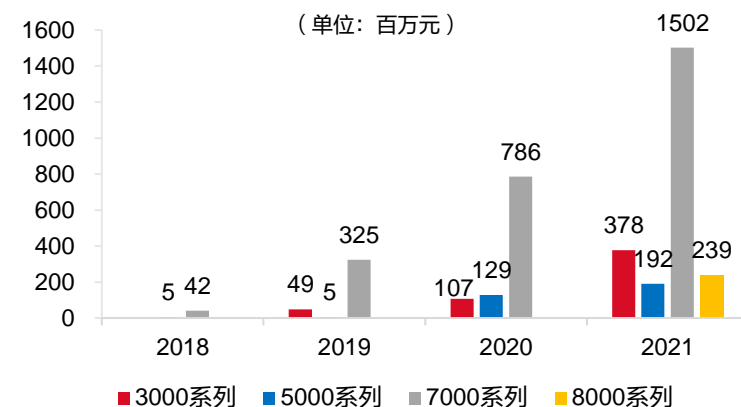
营收、净利润稳定爬坡



毛利率保持稳定，净利率逐年上升



7000系列CPU为营收主要组成部分，2021年占比超65%



注：2022年开始，公司仅拆分为高端处理器和技术服务，2023年上半年高端处理器占比近100%。

资料来源：iFind，东海证券研究所

3.1、海光信息：深算三号研发进展顺利，适配国内外主流AI大模型

- 海光三号、深算二号为公司主力产品，海光四、五号、深算三号研发顺利。海光CPU系列产品海光三号为主力销售产品，海光四号、海光五号处于研发阶段；海光DCU系列产品深算一号为公司GPGPU主要在售产品；深算二号已经发布并实现商用，深算二号实现了在大数据处理、人工智能、商业计算等领域的商业化应用，具有全精度浮点数据和各种常见整型数据计算能力，性能相对于深算一号实现了翻倍的增长；深算三号研发进展顺利。
- 深算系列适配国内外主流大模型，CPU和DCU产品下游应用广泛，得到客户的普遍认可。1) 海光DCU主要面向大数据处理、商业计算等计算密集型应用领域以及人工智能、泛人工智能应用领域展开商用，对文心一言等大多数国内外主流大模型适配良好。依托DCU可以实现LLaMa、GPT、Bloom、ChatGLM、悟道、紫东太初等为代表的大模型的全面应用，达到国内领先水平。在互联网领域，公司的DCU产品已得到百度、阿里等互联网企业的认证，并推出联合方案，打造全国产软硬件一体全栈AI基础设施。2) 海光高端处理器产品已经得到了国内行业用户的广泛认可，逐步开拓了浪潮、联想、新华三、同方等国内知名服务器厂商，开发了多款基于海光处理器的服务器。目前已经应用到了电信、金融、互联网、教育、交通等行业。

海光二代、三代部分CPU产品参数

系列	型号	CPU核心数量	线程数量	典型功耗	最高加速频	Pcle	最高内存频率	内存类型	
海光二代	3000系列	海光3250	8	16	90W	3.0GHz	Pcle 3.0*32	2666MHz	DDR4
	3000系列	海光3330	4	8	35W	3.3GHz	Pcle 4.0*32	3200MHz	DDR4
		海光3350	8	16	65W	3.3GHz	Pcle 4.0*32	3200MHz	DDR4
海光三代	5000系列	海光5380	16	32	70W	3.0GHz	Pcle 4.0*64	3200MHz	DDR4
		海光5390	16	32	95W	3.2GHz	Pcle 4.0*64	3200MHz	DDR4
	7000系列	海光7360	24	48	125W	3.0GHz	Pcle 4.0*128	3200MHz	DDR4
		海光7375	32	64	140W	3.0GHz	Pcle 4.0*128	3200MHz	DDR4
		海光7380	32	64	140W	3.0GHz	Pcle 4.0*128	3200MHz	DDR4
		海光7390	32	64	110W	3.3GHz	Pcle 4.0*128	3200MHz	DDR4

资料来源：公司官网，东海证券研究所

3.2、寒武纪：AI芯片国内领先，云端、边缘、IP授权及软件三位协同发展

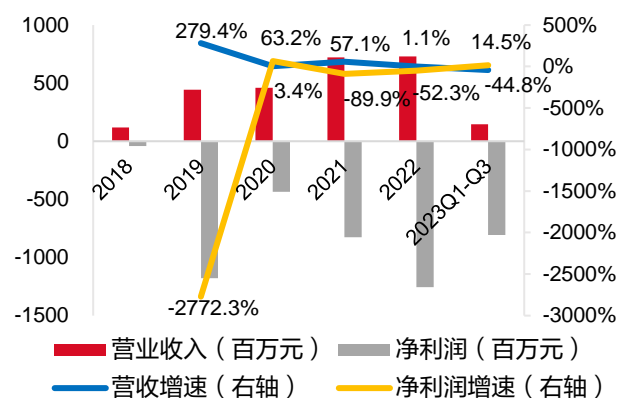
寒武纪成立于2016年，2020年于上交所科创板上市(688256.SH)。主要从事应用于各类云服务器、边缘计算设备、终端设备中人工智能核心芯片的研发、设计和销售。主要产品包括云端产品线、边缘产品线、处理器IP授权及软件，广泛应用于服务器厂商和产业公司，面向互联网、金融、交通、能源、电力和制造等领域的复杂AI应用场景提供充裕算力，推动AI赋能产业升级。

寒武纪主要产品与性能

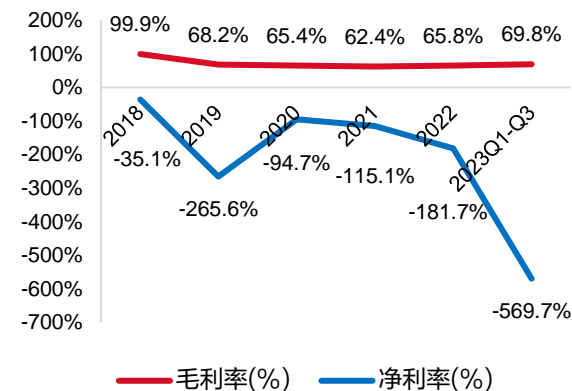
产品线	产品类型	寒武纪主要产品	相关性能	推出时间
云端产品线	云端智能芯片及加速卡	思元100 (MLU100) 芯片及云端智能加速卡	-	2018年
		思元270 (MLU270) 芯片及云端智能加速卡	计算精度支持: INT16,INT8,INT4,FP32,FP16 峰值算力: 128TOPS(INT8); 256TOPS(INT4); 64TOPS(INT16) 内存容量: 16GB DDR4, ECC	2019年
		思元290 (MLU290) 芯片及云端智能加速卡	制程: 7nm 峰值算力: 自适应精度训练算力 512TOPS(INT8);256TOPS(INT16);64TOPS(CINT32) 内存容量: 32GB HBM2高带宽内存	2020年
		思元370 (MLU370) 芯片及云端智能加速卡 (以MLU370-X8为例)	制程: 7nm 计算精度支持: FP32,FP16,BF16,INT16,INT8,INT4 峰值算力: 256TOPS(INT8);128TOPS(INT16); 96TFLOPS(FP16);96TFLOPS(BP16);24TFLOPS(FP32) 内存容量: 48GB LPDDR5	2021年、2022年
训练整机	训练整机	玄思1000智能加速器	峰值算力: 自适应精度算力 2.05 PetaOPS (INT8);1 PetaOPS (INT16);256 TOPS (CINT32) 内存容量: 128GB	2020年
		玄思1001智能加速器	-	2022年
边缘产品线	边缘智能芯片及加速卡	思元220 (MLU220) 芯片及边缘智能加速卡	-	2019年
		寒武纪1A处理器	-	2016年
IP授权及软件	终端智能处理器IP	寒武纪1H处理器	较初代产品其能效比有着数倍提升，广泛应用于计算机视觉、语音识别、自然语言处理等人工智能处理关键领域	2017年
		寒武纪1M处理器	具备了更优性能、更低功耗和更强的完备性，混合支持fp32/fp16/int32/int16/int8/int4位宽，增加了压缩解压缩模块。在上代基础上，可支持个性化人工智能应用，也可用于多路视频实时处理和自动驾驶等领域。	2018年
基础系统软件平台	基础系统软件平台	寒武纪基础软件开发平台 (适用于公司所有芯片与处理器产品)	-	持续研发和升级，以适配新的芯片

资料来源：公司公告，公司官网，东海证券研究所

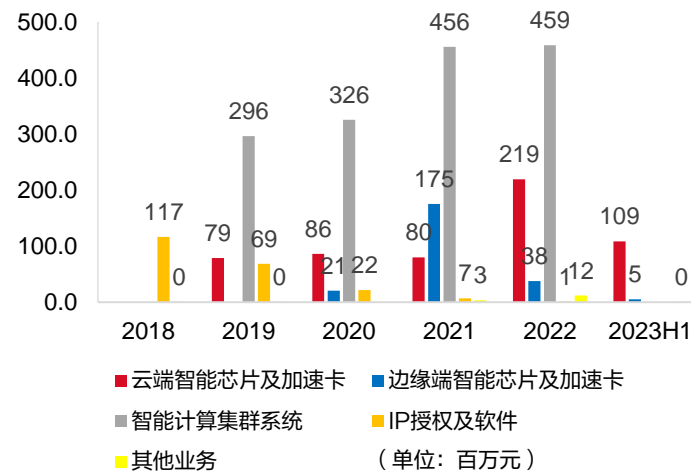
营收短期承压，净利润降幅收窄



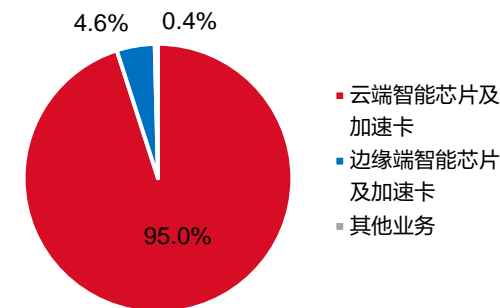
毛利率保持稳定



云端产品线与智能计算集群系统为营收主要组成部分



2023年上半年云端智能芯片及加速卡营收占比95%



资料来源：iFind，东海证券研究所

3.2、寒武纪：互联网、金融、通信等下游领域多点开花

- 公司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。在互联网行业，公司的芯片及加速卡与数家头部互联网企业在视觉、语音、图文识别、自然语言处理等场景下进入了批量销售环节。在金融行业，公司在大语言模型领域与头部银行、头部ISV积极推动技术合作和深度算法适配，为后续的产品大规模落地打下了坚实基础。在通信运营商行业，公司持续在大语言模型应用以及大型集群架构设计上进行探讨和进一步验证性测试工作。

寒武纪在AI领域的业务进展

业务	进展	亮点
智能处理器微架构及指令集	自主研发了五代智能处理器微架构、五代商用智能处理器指令集；第六代智能处理器微架构和指令集正在研发中。	新一代智能处理器微架构及指令集将对自然语言处理大模型和推荐系统的训练推理等场景进行重点优化，将在编程灵活性、能效、功耗、面积等方面提升产品竞争力。
推理软件平台	研发了大模型和AIGC推理业务所需基础软件	<p>(1) 模型性能优化方面，针对语音合成领域、搜索推荐领域、视觉处理领域中新涌现的高频使用网络进行了优化，性能上达到了业务落地要求；</p> <p>(2) 大模型和AIGC推理业务支持方面，研发了大语言模型分布式推理加速库BangTransformer，进行了LLaMA、GLM、BLOOM、GPT-2等主流生成式大语言模型的适配工作；</p> <p>(3) 推理性能优化方面，BangTransformer支持算子融合、张量并行、量化推理、Flash Attention等优化特性，与传统框架运行方式相比有了较大提升；</p> <p>(4) 图像生成领域，基于MagicMind支持了Stable Diffusion等客户需求网络的优化和加速，促进了产品在图像生成领域的业务落地。</p>
基础系统软件平台	持续推进训练软件栈的研发和改进，以客户需求牵引新增功能和通用性支持，并大力推进大模型及推荐系统业务的支持和优化。	<p>(1) 新增功能方面，提供了方便客户模型迁移的多个工具，平滑支持了PyTorch_Lightning等第三方Python库。重点投入了分布式和大规模集群的软件栈支撑，增加支持了DeepSpeed、Megatron、Tutel等分布式训练库。软件栈在加大与社区生态融合的同时也完成了对典型操作系统的支持和发布，支持了生态的发展；</p> <p>(2) 通用性方面，增加了框架算子的支持数量并落实了框架版本升级的规划，完成了重点客户提出的大量定制化算子扩展需求，支撑了多个重点客户的业务落地；</p> <p>(3) 性能方面，通过算子层面大张量支持等新功能的开发，充分发挥了公司硬件产品的架构优势，使网络能够支持高效处理能力以及更大规模的大语言模型。通过重点客户性能需求的牵引，针对多个网络进行了极致优化，分析并实现了多个算子和框架层面的性能优化点，并在通信上持续进行了单机多卡的低延迟优化；</p> <p>(4) 大模型和AIGC训练领域，完善了大模型的训练软件栈研发，进行了GLM模型、LLaMA模型、GPT模型等模型的微调及预训练支持工作，验证了张量并行、流水并行、序列并行等并行技术，达到业务落地的精度和性能要求。</p>

资料来源：公司公告，东海证券研究所

3.3、澜起科技：互联类芯片业务筑基，服务器平台产品持续放量

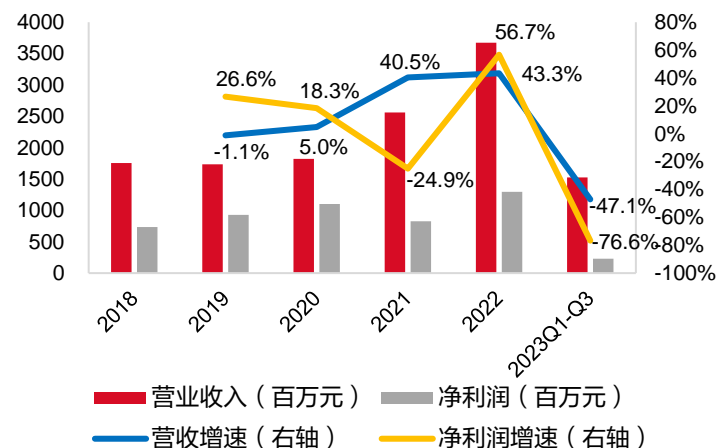
澜起科技成立于2004年，2019年于上交所科创板上市(688008.SH)。公司为云计算和AI领域提供高性能、低功耗的芯片解决方案，拥有互连类芯片和津速服务器平台两大产品线。互连类芯片产品主要包括内存接口芯片、内存模组配套芯片、PCIe Retimer芯片、MXC芯片、CKD芯片等，津速服务器平台产品包括津速CPU和混合安全内存模组。同时，公司正在研发基于“近内存计算架构”的AI芯片。

澜起科技主要产品与性能

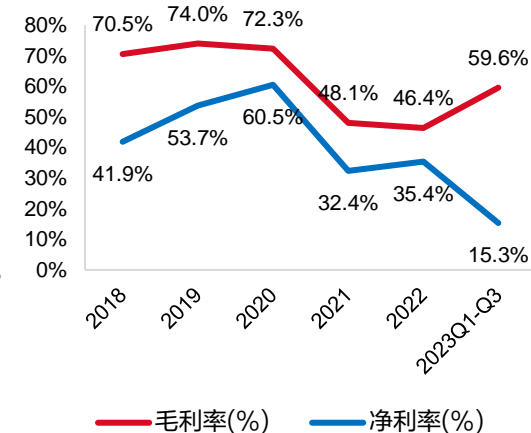
产品	细分分类	产品类型	性能/特点	应用
互连类芯片	内存接口芯片	DDR3/2 (以DDR3 RCD为例)	最高速率1866Mbps 工作电压1.5/1.35/1.25V	DDR3 RDIMM
		DDR4 (以Gen2 Plus DDR4 RCD为例)	最高速率3200Mbps 工作电压1.2V	DDR4 RDIMM, LRDIMM和NVDIMM
		DDR5 (以Gen3 DDR5 RCD为例)	最高速率6400Mbps	DDR5 RDIMM
	内存模组配套芯片	DDR5 PMIC, DDR5 SPD Hub, DDR5 TS	-	-
	MRCD/MDB芯片	-	-	-
	PCIe Retimer芯片	16通道PCIe® 5.0/CXL® 2.0 Retimer	354-ball FCCSP封装	服务器、存储设备 和硬件加速器
互联类芯片	MXC芯片	8通道PCIe® 4.0 Retimer	332-ball FCCSP封装	内存AIC扩展卡; EDSFF内存模组
		16通道PCIe® 4.0 Retimer	354-ball FCCSP封装	
	CKD芯片	Type 3 CXL®内存扩展控制器芯片	CXL标准: CXL® 1.1/2.0;CXL® x8 lanes DDR标准: JEDEC DDR4/DDR5 封装: 767球 FCCSP	-
津速®服务器平台产品线	津速®CPU	第四代津速®CPU	最大核心数: 48核 最大线程数: 96线程 最大共享缓存: 105MB 最高睿频频率: 4.2GHz	-
		第五代津速®CPU	最大核心数: 48核 最大线程数: 96线程 最大共享缓存: 260MB 综合浮点运算性能提升: 40%	-
	混合安全内存模组 (HSDIMM®)	-	-	-
AI芯片	-	-	-	-

资料来源：公司公告，公司官网，东海证券研究所

营收短期承压，净利润同比转负

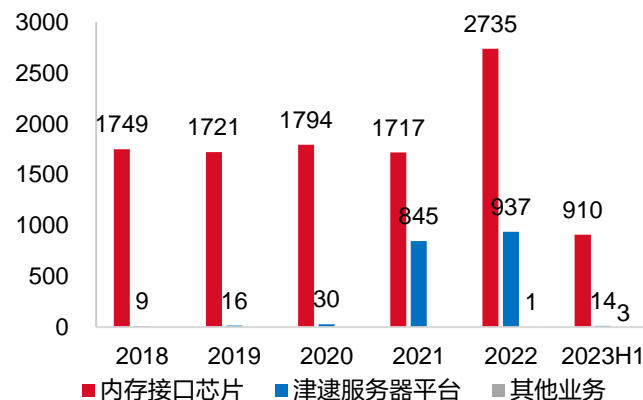


毛利率上浮显著



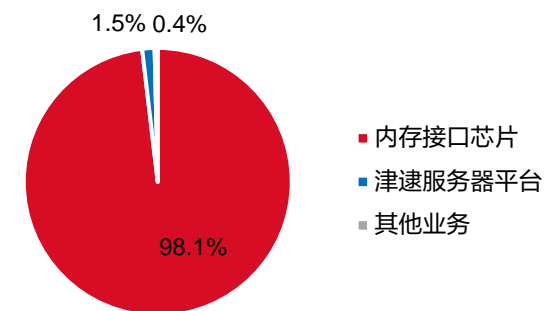
津速服务器平台业务逐步放量

(单位：百万元)



资料来源：iFind，东海证券研究所

2023年上半年内存接口芯片营收占比超过98%



3.3、澜起科技：DDR5芯片迭代更新，第五代津逮CPU放量在即

- 公司凭借具有自主知识产权的高速、低功耗技术，致力于为新一代服务器平台提供符合JEDEC标准的高性能内存接口解决方案，目前DDR5芯片产品已推出至第四子代。随着JEDEC标准和内存技术的发展演变，公司先后推出了DDR2-DDR5系列内存接口芯片。2024年1月4日，公司推出DDR5第四子代寄存时钟驱动器芯片(DDR5 RCD04)，该芯片支持高达7200MT/s的数据速率，较DDR5第一子代RCD速率提升50%，以应对新一代服务器平台对内存速率和带宽不断攀升的需求。
- 第五代津逮CPU已发布，性能大幅提升。2023年12月18日，澜起科技发布其全新第五代津逮CPU，相比第四代，其单颗CPU最高支持48个核心、96个线程，最大三级缓存容量达260MB；支持的DDR5内存速度最高达5600MT/s，CPU之间互连的UPI速度最高达20GT/s；基于LINPACK测试，其综合浮点计算性能最高提升近40%。

澜起科技各业务研发进展

产品	主要进展
内存接口芯片	2023年上半年，公司推进DDR5第二子代RCD芯片量产准备工作。 2023年10月，公司在业界率先试产DDR5第三子代RCD芯片。根据主流CPU厂商公布的最新产品路线图，其支持内存速率6400MT/S的新一代服务器CPU平台计划于2024年发布，该芯片预计将跟随该CPU平台的发布而开始规模出货。 2024年1月，公司推出DDR5第四子代RCD芯片，该芯片支持高达7200MT/S的数据速率。目前公司已将该产品工程样片送样给主要内存厂商。
MRCB/MDB芯片	2023年上半年，基于客户对DDR5第一子代MRCB/MDB芯片工程样片的反馈意见，公司推进量产版本的研发。根据主流CPU厂商公布的最新产品路线图，其支持MCRDIMM的新一代服务器CPU平台计划于2024年发布，该芯片预计将跟随该CPU平台的发布而开始规模出货。
PCIe Retimer芯片	2023年1月，公司的PCIe 5.0/CXL 2.0 Retimer芯片实现量产，有望在2024年实现规模出货。
MXC芯片	2023年上半年，公司完成了CXL内存扩展控制器芯片（MXC）量产版本的流片及样品制备，正在推进量产前的质量认证及客户认证等相关工作。
CKD芯片	2023年上半年，公司完成了业界首款DDR5第一子代时钟驱动器（CKD）量产版本的流片及样品制备，正在推进量产前的质量认证及客户认证等相关工作。
津逮®CPU	2023年1月，公司正式发布第四代津逮®CPU产品。 2023年12月18日，澜起科技正式向外界发布其全新第五代津逮®CPU。
AI芯片	2023年上半年，公司开展了第一代AI芯片工程样片的相关测试及验证工作，并在相关应用平台进行业务适配。2023年上半年，公司新申请AI芯片相关发明专利8项。截至2023年6月30日，公司累计申请AI芯片相关发明专利46项，其中4项已获授权。

资料来源：公司公告，公司官网，东海证券研究所

3.4、中际旭创：全球光模块龙头厂商，中低、高速光模块和光组件全覆盖

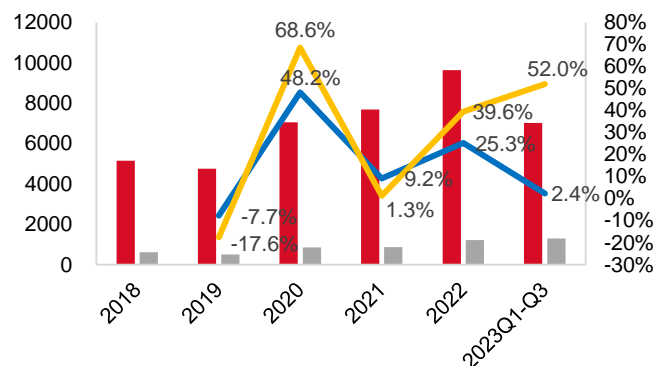
- **公司简介：**中际旭创(300308.SZ)是专业的高速光模块解决方案提供商，公司主要从事高端光通信收发模块以及光器件的研发、设计、封装、测试及销售，主要产品为中低速光通信模块、高速光通信模块、光组件。
- **主要业务：**其全资子公司苏州旭创致力于高端光通信收发模块的研发、设计、封装、测试和销售；控股子公司成都储翰专注接入网光模块和光组件生产及销售。在Lightcounting发布的2022年度光模块厂商排名中，中际旭创和Coherent并列全球第一。

中际旭创部分光模块产品

产品系列	产品特性	符合标准	应用场景
800G OSFP	拥有全面的800G OSFP光模块产品组合，包括4x100Gx2和8x100G两种架构方案，除了传统的EML设计，还采取了以硅光为基础的方案来满足短距离传输需求	符合IEEE802.3ck和OSFP MSA标准，并支持CMIS4.0	主要应用于800G以太网、数据中心和云网络
800G QSFP-DD	拥有全面的800G QSFP-DD光模块产品组合，包括4x100Gx2和8x100G两种架构方案，除了传统的EML设计，还采取了以硅光为基础的方案来满足短距离传输需求	符合IEEE802.3ck和QSFP-DD 800 MSA标准，并支持CMIS4.0	主要应用于800G以太网、数据中心和云网络
400G QSFP-DD	拥有全面的400G QSFP-DD光模块产品组合	符合IEEE 802.3bs 和QSFP-DD MSA标准	主要应用于400G以太网、数据中心和云网络
400G OSFP	拥有全面的400G OSFP光通信模块产品组合，包括4x50Gx2和4x100G两种架构方案	符合IEEE 802.3bs 和OSFP MSA标准	主要应用于400G以太网、数据中心和云网络
100G QSFP28 Single Lambda	具有小型化、低功耗和高速率的特点	符合IEEE 802.3bm, IEEE 802.3cd和QSFP28 MSA标准	主要应用于100G以太网
100G QSFP+	包括SR4, SR4 CPRI, AOC, AOC 100G-4x25G, CWDM4, eCWDM4, eCWDM4 ET PSM4, PSM4 pigtail, LR4 Ethernet和ER4 Lite系列，该系列产品采用LC或MPO光口；具有功耗低、体积小、速率高等特性，有利于数据中心增加容量、提高端口密度和降低功耗	兼容IEEE802.3bm, SFF-8636等标准	主要应用于100G数据中心内部网络、数据中心互联、城域网等环境，也可应用于5G无线网络
40G QSFP+	包括SR4, eSR4, IR4, LR4, ER4, LX4, PSM IR4, PSM LR4, AOC and AOC breakout系列。该系列产品采用LC或MPO光口；具有功耗低、体积小、速率高等特性，有利于数据中心增加容量、提高端口密度和降低功耗	兼容IEEE802.3bm, SFF-8436等标准	主要应用于大型数据中心、园区网络、城域网等环境
25G SFP28	包括SR, AOC, LR, ER商业温度系列，以及LR, BiDi, CWDM, LWDM, ER等工业温度系列。这些产品采用LC光口；具有功耗低、体积小、速率高、宽温度范围等特性	兼容IEEE802.3by, SFF-8472等标准	主要应用于数据中心、5G网络、25G以太网、光纤通道等环境
10G SFP+ SONET	拥有全面的SONET系列产品，包括LR, ER, ZR, DWDM ER, DWDM ZR系列	符合SONET OC192/SDH STM64与IEEE802.3ae标准	主要应用于SONET(OC-192)/SDH(STM64)传输网络环境
10G SFP+ Ethernet	包括LR, ER, ZR和DWDM (40km与80km)系列，该系列产品采用LC光口；具有功耗低、体积小、速率高等特性	兼容IEEE802.3ae, SFF-8472, SFF-8431等标准	主要应用于数据中心、城域网、无线网络、传输网络等环境

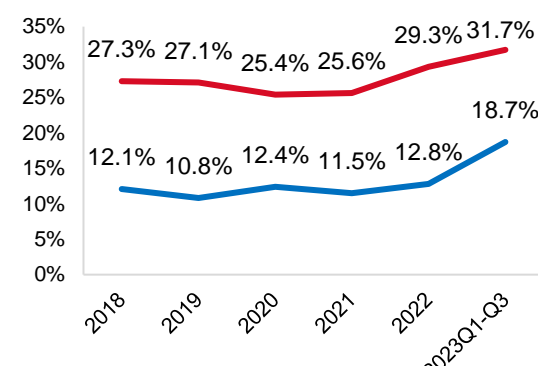
资料来源：公司公告，东海证券研究所

营收增速放缓，净利润持续高速增长



■ 营业收入 (百万元) ■ 净利润 (百万元)
— 营收增速 (右轴) — 净利润增速 (右轴)

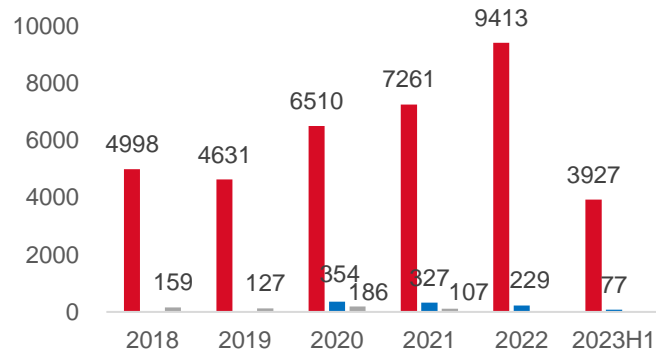
毛利率逐步上升



— 毛利率 (%) — 净利率 (%)

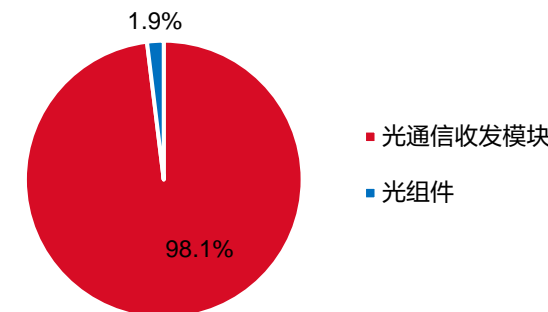
光通信收发模块业务稳步扩张

(单位: 百万元)



■ 光通信收发模块 ■ 光组件 ■ 其他

2023年上半年光通信收发模块业务营收占比超过98%



资料来源：iFind，东海证券研究所

3.4、中际旭创：800G光模块业务稳步爬坡，1.6T光模块进展国内领先

- 公司的光模块产品以技术优良、性能稳定、供应可靠等特性获得了下游客户的认可，与全球领先的云数据中心客户和国内外主流通信设备厂商形成了长期稳定的合作关系。公司为云数据中心客户提供100G、200G、400G和800G的高速光模块，为电信设备商客户提供5G前传、中传和回传光模块，应用于城域网、骨干网和核心网传输光模块以及应用于固网FTTX光纤接入的光器件等高端整体解决方案。
- AI大模型催生了公司800G及以上高速光模块业务高增长。2023年以来，随着ChatGPT为代表的生成式人工智能大语言模型的发布，催生了AI算力需求的激增，进而拉动了800G光模块需求的显著增长，并加速了光模块向800G及以上产品的迭代升级，公司800G等高端产品取得了良好的订单和市场份额，同时1.6T光模块产品预计于2024年下半年有相关重点客户导入。

中际旭创业务进展

业务	产品	最新进展
光通信领域	800G光模块	800G光模块业务从2023年第二季度开始开始起量，几乎每个季度都保持了订单和出货量的环比增长，该业务基本围绕和AI服务器相配套的需求，且该需求基本不会受到1.6T光模块上量的影响（部分客户2025年才开始正式部署800G）。
	1.6T光模块	预计AI重点客户最快在2024年下半年开始采购和部署1.6T光模块，相关产品预计在2025年开始规模上量。
	400G、800G硅光模块	400G和800G已有一些型号采用了硅光方案并通过了大客户的认证，已开始出货，后续预计大客户会加大对硅光模块的采购比例，预计2024年400G和800G的硅光模块都有机会进一步放量和扩大出货比例。
汽车光电子	车载光互联、新一代激光雷达等	2023年上半年，公司持续推进产业投资和多元化战略，设立了全资子公司江苏智驰网联控股有限公司，同时收购重庆君歌电子科技有限公司62.45%的股权，从而实现公司在汽车光电子领域新的业务拓展，并将公司在光通信领域的技术储备，通过君歌电子的市场渠道和客户资源加快推进如车载光互联、新一代激光雷达等汽车光电子新产品客户导入与量产，实现较好的产业协同效应。

资料来源：公司公告，东海证券研究所

3.5、天孚通信：提供光器件产品和一站式解决方案的平台型国内龙头企业

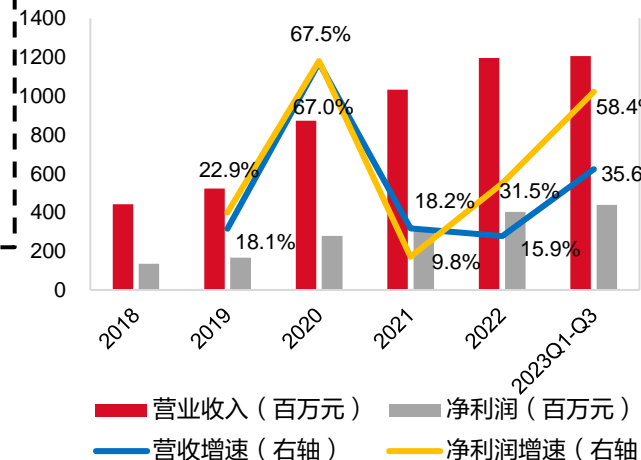
- **公司简介：**苏州天孚光通信股份有限公司2005年成立，2015年登陆深交所创业板(300308.SZ)，是业界领先的光器件整体解决方案提供商和先进光学封装制造服务商。
- **公司业务：**产品包括光无源器件、光有源器件，广泛应用于光纤通信、光学传感、激光雷达、生物光子学等领域。公司目前已发展了十三大产品线，八大方案和针对激光雷达、医疗监测板块的光学类器件，为拥有多种器件和封装技术能力的复合平台型企业。

天孚通信八大光器件解决方案

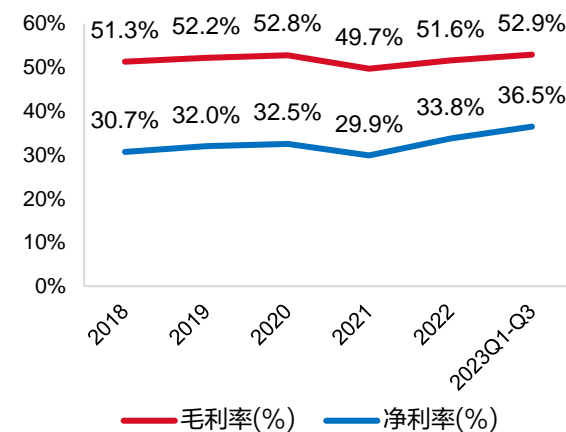
光器件解决方案	解决方案示意图	优势	主要应用领域
高速同轴光器件产品解决方案		具有同轴TOSA/BOSA(ODM/OEM)定制高精度自动对准和封装能力； 高精度、高可靠性贴装和金丝键合能力； TO-CAN封装高精度、高可靠性；	电信通信、 数据中心、企业网
高速光引擎/BOX器件封装解决方案		具备AWG/TFF封装高精度、高可靠性贴装和金丝键合能力； 具有光学模拟和分析能力，能够提供各种定制的带LENS组件； 高精度芯片贴装技术能力； 高分辨率3D视觉影像自动耦合能力；	电信通信、 数据中心
微光学产品解决方案		TFC拥有微光学光路模拟/设计/装配，光学镀膜设计和加工能力，能根据客户要求设计进行光学镀膜； 拥有高精度铁和技术，可提供稳定TFF贴合； 拥有光学模拟分析能力，可以定制设计和加工类型的组件带透镜； 拥有各种类型自由空间隔离器设计和隔离器芯片与插芯贴合能力	电信通信、 数据中心
波分复用(AWG)产品解决方案		拥有硅，二氧化硅等光学材料加工能力，可根据客户要求定制加工； 拥有玻璃切割，FA加工研磨能力，可根据客户要求定制各类型产品； 拥有光学模拟分析能力和自动化开发能力，拥有自动光纤对准系统； 拥有高速同轴Mini型器件耦合组装能力。	数据中心
PSM/DR系列光器件无源产品解决方案		TFC拥有高精度玻璃光纤阵列(FAU)设计，加工和组装能力； 拥有COB(45度，42.5度)FAU凸纤研磨能力； 拥有FA表面光学镀膜设计和加工能力； 拥有高精度隔离器与FAU的贴合工艺	数据中心
PM保偏+FAU无源光器件产品解决方案		低插入损耗； 高消光比； 高回波损耗； 良好的环境稳定性和可靠性。	电信通信、 数据中心
SR&OBO用塑料透镜和光阵列产品解决方案		具有光学设计、高精度成型和注塑COB塑料透镜阵列组件的能力； 具有100G/200G/400G SR收发器短线组件定制设计能力； 具有光纤激光切割技术的的能力； 具有高精度COB封装能力。	数据中心
AOC系列无源光器件产品解决方案		TFC可配合客户进行定制化的100G/200G/400G AOC系列光连接解决方案设计； 拥有各种类型陶瓷插芯，MT插芯及非标塑料插芯设计和生产能力； 拥有配套金属加工件及注塑件自主设计和加工能力； 拥有MT高回损加工工艺，RL>40dB。	数据中心

资料来源：公司公告，公司官网，东海证券研究所

营收、净利润增速上行显著

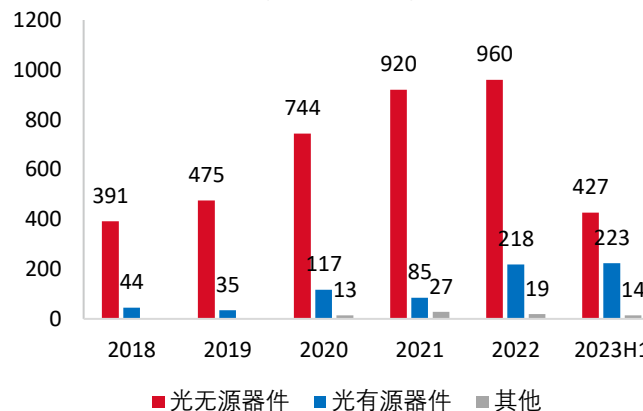


毛利率、净利率保持稳定

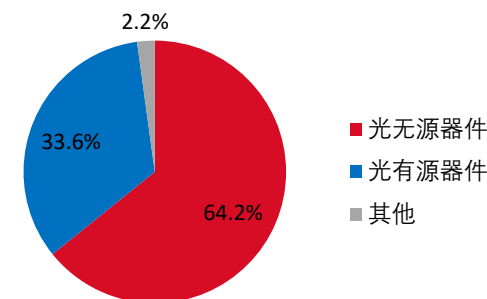


光无源器件、光有源器件业务迅速扩张

(单位：百万元)



2023年上半年光无源器件业务营收占比超过60%



资料来源：iFind，东海证券研究所

3.5、天孚通信：AI持续拉动需求，高速光引擎业务构筑长期成长曲线

- **AI技术迭代升级带动对算力网络的高增长需求，同步推动公司高速光器件市场持续增长。**全球数据中心建设对光器件产品需求的持续稳定增长，公司为400G、800G等高速光模块提供一站式产品解决方案。2023年，依托行业向好的发展趋势，公司稳步推进高速光引擎募投项目建设，同时加速高速光器件的研发和规模量产，最大化满足客户需求，取得阶段性成效。
- **公司具备多个光模块技术与先进光学技术。**2024年2月27日，公司的十六条产品线&八大解决方案，覆盖了所有光模块用的光无源器件，可以为客户提供多技术平台、多应用场景的光器件整体解决方案。其次是先进光学封装制造，已经完成了OSA、BOX、COB、TO、硅光等多种光器件封装平台的光引擎、高速光器件的ODM/OEM业务。
- **公司具备1.6T/800G光模块配套业务。**公司宣布将于2024年3月26日至28日在美国加州圣地亚哥会展中心举办的第49届光网络与通信研讨会及博览会（OFC2024）重点展示为1.6T/800G光模块配套应用的Mux TOSA、Demux POSA、Lensed FAU等光引擎产品和解决方案。

天孚通信项目研发投入情况

主要研发项目名称	项目目的	项目进展	拟达到的目标	预计对公司未来发展的影响
800G光器件开发	开发并量产下一代数据中心用光引擎	量产	满足客户定制化需求	增强公司在前沿产品的市场竞争力
光引擎研发	垂直整合公司既有有源&无源产品线，为下游客户提供一站式解决方案	部分产品量产；部分产品开发中	持续夯实光引擎平台，持续为客户提供高性价比的解决方案	有源、无源双擎轮动，加强公司对前沿技术的理解和布局，灵活的ODM/OEM开发模式，持续保持公司的竞争力。
单波100G光器件开发	垂直整合公司有源&无源产品线优势，给客户最佳解决方案	小批量	增强同轴类TO封装平台的持续领先优势	单波100G与400G对传，有望持续放量，获得较高销售额
车载激光雷达用光器件的开发	基于公司领先的无源&有源平台，为雷达客户提供高性价比的解决方案	小批量	扩展公司产品线至汽车行业	打造新的增长点，拓宽公司业务领域
激光芯片集成高速光引擎研发	基于公司的垂直整合平台，开发适用于下一代数据中心用光引擎	小批量	开发CPO用光引擎，为客户提供一站式解决方案	持续保持技术领先，巩固与头部客户的持续合作
保偏光器件的研发	开发高速光模块用零组件，集成隔离器、FA、连接器件、组件等无源器件，提高产品的小型化及集成度	部分客户已批量交付，部分产品小批量验证中	批量应用于相干光器件	增强公司在前沿产品的市场竞争力
小型一体化组件的开发	在组件基础之上，开发多套一小型化，一体式组件，以满足客户高密度需求项目上对组件的应用需求	小批量交付阶段	满足客户定制化需求	打造新的增长点，拓宽公司业务领域

资料来源：公司公告，东海证券研究所

3.6、新易盛：专注点对点光模块，国内少数拥有800G批量交付能力厂商

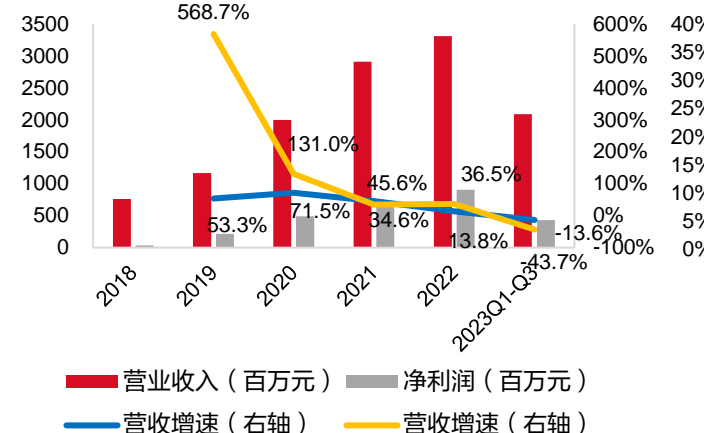
- **公司简介：**成都新易盛通信技术股份有限公司成立于2008年，2016年于深交所创业板上市(300502.SZ)。公司从事高性能光模块的研发、生产和销售。
- **公司业务：**产品服务于数据中心、数据通信、5G无线网络、电信传输、固网接入、智能电网、安防监控等领域的国内外客户。细分来看，主要为云数据中心客户提供100G、200G、400G、800G产品；为电信设备商客户提供5G前传、中传和回传光模块以及应用于城域网、骨干网和核心网传输的光模块；为智能电网和安防监控网络服务商提供光模块解决方案。

新易盛1.6T、部分800G、400G光模块产品介绍

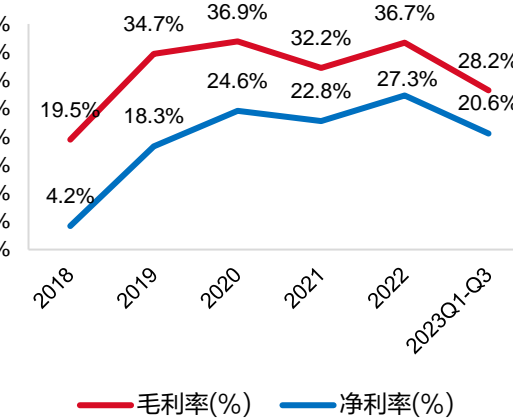
光模块产品系列	产品照片	产品简介	主要应用场景
OSFP-XD 1.6T		符合OSFP-XD MSAs的最新版本；固件支持CMIS 5.0和更新版本；涵盖 DR8, 2xFR4和4xFR2传输接口。	数据中心、1.6T 以太网、云计算网络等
QSFP-DD 800G 单波200G		符合QSFP-DD800 MSA的最新版本；固件支持CMIS 5.0和更新版本；涵盖DR4+、1xDR4、1xFR4和2xFR2传输接口。	
OSFP 800G 单波200G		符合最新版本的OSFP MSA的最新版本；固件支持CMIS 5.0和更新版本；涵盖DR4+、1xDR4、1xFR4和2xFR2传输接口。	数据中心、800G 以太网、云计算网络等
QSFP-DD 800G 单波100G		符合QSFP-DD800 MSA的最新版本；固件支持CMIS 4.0和更新版本；涵盖SR4.2、SR8、DR8、2xFR4和2xLR4传输接口，新推出800G BIDI、800G LPO和800G低功耗产品。	
OSFP 800G 单波100G		符合最新版本的OSFP MSA的最新版本；固件支持CMIS 4.0和更新的版本；涵盖SR4.2、SR8、DR8、2xFR4和2xLR4传输接口，新推出800G BIDI、800G LPO和800G低功耗产品。	
QSFP112 400G		符合QSFP112 MSA的最新要求，固件支持CMIS 4.0或更新版本，支持SR4、DR4、FR4和LR4传输接口，可满足超低功耗要求。	
QSFP-DD 400G		符合QSFP-DD MSA的最新要求，固件支持CMIS 4.0或更新版本，支持SR8、DR4、FR4和LR4传输接口，可满足超低功耗要求。	数据中心、400G 以太网、云计算网络等
OSFP 400G		符合OSFP MSA的最新要求，固件支持CMIS 4.0或更新版本，支持SR8、DR4、FR4和LR4传输接口，可满足超低功耗要求。	

资料来源：公司公告，公司官网，东海证券研究所

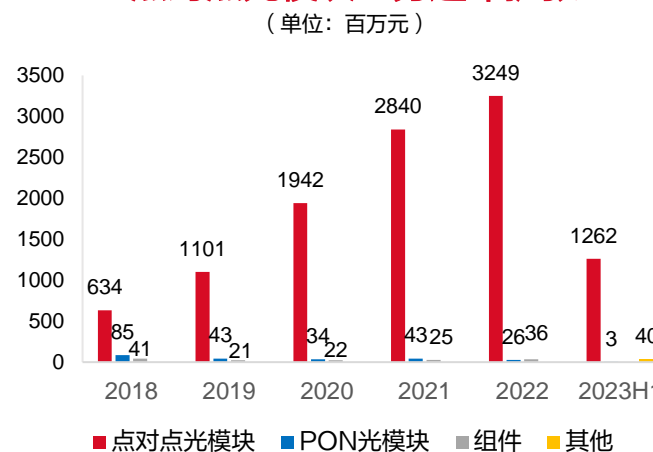
营收、净利润短期承压



毛利率、净利率

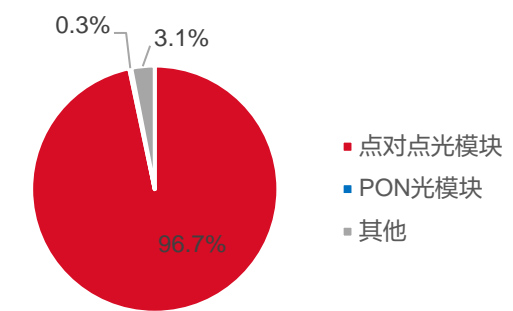


点对点光模块业务逐年扩张



资料来源：iFind，东海证券研究所

2023年上半年点对点光模块业务营收占比超过95%



3.6、新易盛：1.6T业务扬帆启程，硅光、LPO等新技术方案全面布局

- 公司在5G与数通市场布局已久。公司一直专注于光模块的研发、生产和销售，是国内少数批量交付运用于数据中心市场的100G、200G、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，已成功研发出涵盖5G前传、中传、回传的25G、50G、100G、200G系列光模块产品并实现批量交付。
- 公司高速光模块技术适应AI大模型发展需求。把握基于大模型的生成式人工智能AIGC给光模块带来的旺盛的市场需求，持续进行行业新技术、新产品的研究，目前已成功推出800G的系列高速光模块产品，基于硅光解决方案的800G、400G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，同时在OFC2023期间推出了1.6T相关光模块产品，目前正在按计划正常推进中。

新易盛主要业务进展

相关业务	公司布局与最新进展
800G光模块	已成功推出800G的系列高速光模块产品，实现批量出货，公司持续积极推进客户拓展工作，目前正按照计划有序进行，预计2024年行业对800G光模块的需求量较2023年会有同比增长，公司未来具体的放量进度将取决于市场及客户的需求情况
1.6T光模块	公司已在OFC2023期间推出了1.6T相关光模块产品，目前正在按计划正常推进中。800G光模块目前正处于逐渐的商用过程中，1.6T产品未来的进展也与产品的迭代周期相关
LPO方案	LPO方案作为对传统DSP产品方案的补充，公司已在OFC2023期间推出了800G线性驱动可插拔光学器件（LPO）系列产品，公司的800GLPO产品组合包括用于多模和单模应用的模块，目前正在积极推进相关产品的测试和验证工作
硅光技术	公司已在硅光产品方向做了充分的布局，目前在400G相关产品端已实现量产
薄膜铌酸锂方案	公司在相关产品上已有布局，有机会在800G产品阶段实现量产

资料来源：公司公告，东海证券研究所

3.7、光迅科技：拥有从芯片、器件、模块到子系统的垂直集成能力

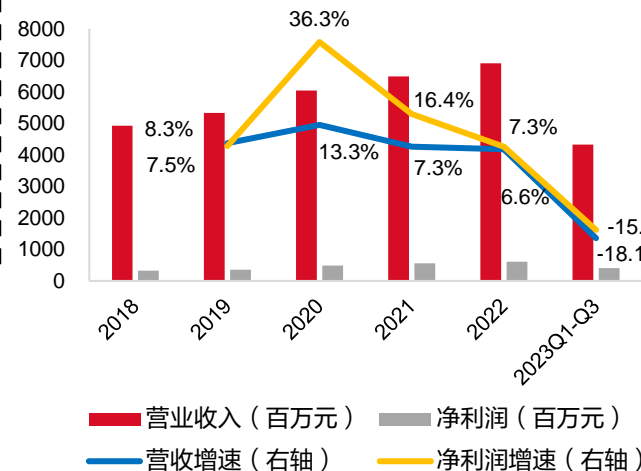
- **公司简介：**武汉光迅科技股份有限公司成立于2001年，主营业务为光电子器件、模块和子系统产品的研发、生产及销售。产品主要应用于电信光通信网络和数据中心网络，可分为传输类产品、接入类产品和数据中心类产品。
- **公司业务：**2022Q2~2023Q1公司在全球光器件行业排名保持第四，在电信传输、数据通信、接入网三大细分市场的全球排名分别为第4、5、3名。公司产品覆盖全面，拥有光芯片、耦合封装、硬件、软件、测试、结构和可靠性七大技术平台，支撑公司有源器件和模块、无源器件和模块产品。

光迅科技主营业务

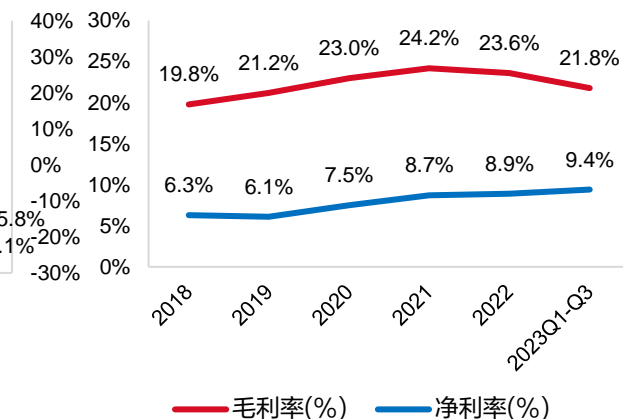
业务	产品分类	细分产品	功能	应用领域
传输类产品 (有源&无源)	传输类光模块	直接调制光模块和相干光模块	支持10Gb/s、100Gb/s、400Gb/s等速率	城域、远距离长途传输系统
	光放大器	EDFA(掺铒光纤放大器)、RFA(拉曼放大器)、RFA/Hybrid(混合光放大器)以及EDFAarray(掺铒光纤放大器阵列)、PluggableAmplifier(可插拔光放大器)等	用于线路两端和中间对光功率进行放大,提升光通信传输距离	
		波长和功率管理类光器件	MUX/DEMUX(复用/解复用)、WSS(波长交换开关)、MCS(MulticastSwitch)等	用于实现合分波、波长交换等功能
	监控类光器件	OTDR(光时域反射仪)、OPM(光功率监测仪)等	用于监控线路中光信号的功率、光信噪比能指标	
数据通信产品	数据中心内光模块		支持100Gb/s、200Gb/s、400Gb/s、800Gb/s等速率,支持QSFP、QSFP-DD、OSFP等封装,支持100m、2km、10km等传输距离	
	数据中心互联光模块	400GZR QSFP-DDDCO、400GZR+ QSFP-DDDCO等		
接入类产品	用于存储网络的光模块	16G FC和32G FC光模块		Fiberchannel存储网络
	固网接入产品	GPONOLT/ONU、10GPON(10GEPON、10G GPON、10G ComboPON)的BOSA和光收发模块等		
	无线接入产品	4GLTE和5G网络用CPRI/eCPRI的前传光收发模块	支持10km、20km、40km等传输距离,支持灰光、CWDM、LWDM、MWDM等波长方案	
其他		10G、100G、400G长跨距、光线路保护、分光放大以及传感类方面的解决方案		

资料来源：公司公告，东海证券研究所

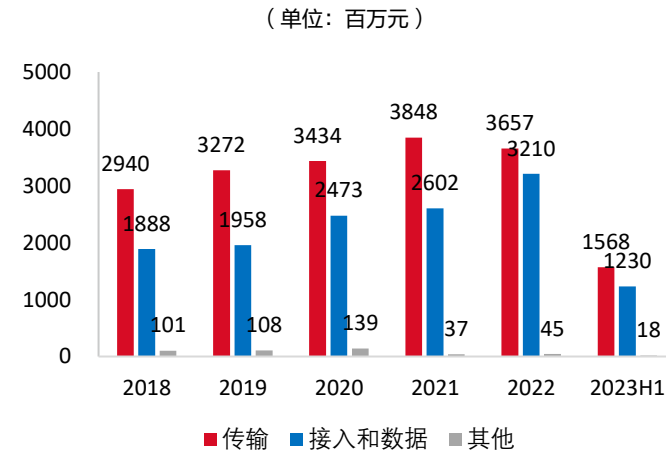
营收、净利润短期承压



毛利保持稳定、净利率稳定爬坡

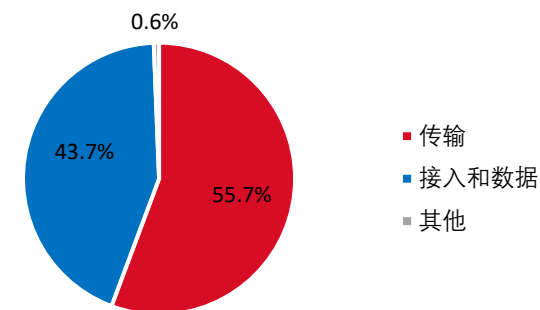


传输、接入和数据三轮驱动



资料来源：iFind，东海证券研究所

2023年上半年传输类业务占比超过55%



3.7、光迅科技：CPO、硅光技术先发布局，800G研发和市场均有突破

- 算力需求爆发拉动光模块业务更新迭代高速发展。数据中心光模块处于从100G向200G、400G的更替阶段，800G产品也逐渐开始使用。公司迅速响应市场需求，在送样测试认证和海外产能建设方面投入较大，公司在马来西亚建立了工厂，目前已经投产。公司800G多模和单模的研发进展都比较顺利，送样后陆续获得了订单。同时，1.6T光模块方面公司已在OFC2023上推出了demo版本。
- 公司于CPO、硅光方面的布局较早，各项技术均有所突破，LPO已进入客户测试阶段。硅光方面从100G、200G开始便进行了相关研发投入，目前进展较为顺利，已进入量产阶段，400G、800G已经开始陆续出货。

光迅科技主要业务进展

业务	相关进展
400G、800G光模块	AI对400G、800G均有较大驱动作用，国内市场以400G光模块为主，国外市场以800G光模块为主。2023年公司400G和800G在研发和市场突破较大，400G批量交货，800G多模和单模的进展都较顺利，在国内国际送样后陆续获得了订单。400G在长距进展顺利，在相干领域取得了较大进展，国内和国外均获得部分订单。
1.6T光模块	还在研发验证阶段，在2023OFC上已给出demo版本。
100GEML和vcSEL	100GEML预计在2024年上量，100GvcSEL还在预研发阶段。
LPO、CPO	CPO方面从2020年开始相关布局研发，光源方面内部已通过研发验证，MPO连接器进行了专利和研发方面的布局。LPO方面，2023年公司开始进行大幅投入，目前有部分芯片和模块在客户进行送样测试，进展较顺利。
硅光技术	相关布局较早，相关研发投入可追溯至100G、200G光模块产品，从400G、800G光模块产品开始可以切入市场。公司目前硅光的进展比较顺利，已进入量产阶段，400G、800G已经开始陆续出货。硅光里应用的LPO公司也在进行生产测试，在硅光400G相干方面公司也有布局。
相干光模块	公司布局超过十年，2023年实现较大突破，在部分客户处通过了认证测试。

资料来源：公司公告，东海证券研究所

3.8、源杰科技：国内领先的IDM平台光芯片供应商

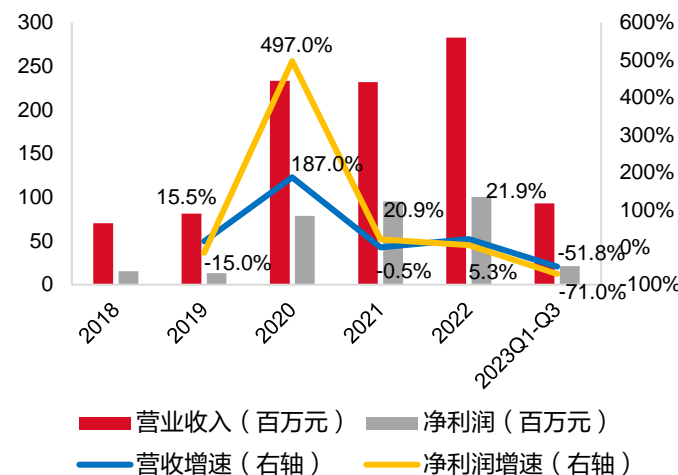
- **公司介绍：**陕西源杰半导体科技股份有限公司成立于2013年，2022年于科创板上市。公司聚焦于光芯片行业，主营业务为光芯片的研发、设计、生产与销售，产品主要包括2.5G、10G、25G、50G及更高速率激光器芯片系列产品等，目前主要应用于电信市场、数据中心市场、车载激光雷达等领域。
- **公司业务：**公司已建立了包含芯片设计、晶圆制造、芯片加工和测试的IDM业务体系，拥有多条覆盖MOCVD外延生长、光栅工艺、光波导制作、金属化工艺、端面镀膜、自动化芯片测试、芯片高频测试、可靠性测试验证等全流程自主可控的生产线，已实现向国际前十大及国内主流光模块厂商批量供货，已成为国内领先的光芯片供应商。

源杰科技主要产品情况

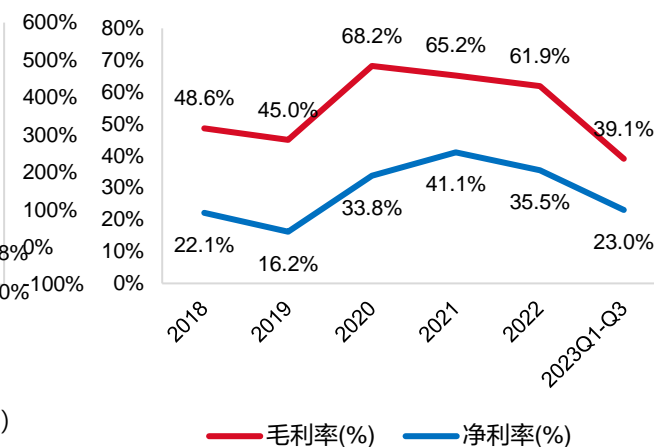
产品速率	产品类型	应用领域
2.5G	1310nm DFB激光器芯片	光纤接入GPON 光纤接入：光纤传输的光通信系统中，光网络单元（ONU）与光线路终端（OLT）之间的光信号传输
	1490nm DFB激光器芯片	
	1270nm DFB激光器芯片	
	1550nm DFB激光器芯片	
10G	1270nm DFB激光器芯片	4G/5G移动通信网络 4G/5G基站：电信运营商通信网络主要包括骨干网与城域网，城域网分为核心层、汇聚层、接入层，其中接入层通常为终端用户连接或访问网络的部分。电信运营商在接入层建设大量通信基站，将用户数据转换为光信号，并通过汇聚层、核心层网络回传至骨干网
	1310nm FP激光器芯片	
	1310nm DFB激光器芯片	
	CWDM 6波段 DFB激光器芯片	
25G	CWDM 6波段 DFB激光器芯片	5G移动通信网络
	LWDM 12波段 DFB激光器芯片	
	MWDM 12波段 DFB激光器芯片	
	CWDM 4波段 DFB激光器芯片	
50G	LWDM 4波段 DFB激光器芯片	数据中心100G 数据中心建设：互联网公司、云计算建设的大型数据中心内部的数据传输、数据中心之间的数据传输
	PAM4 CWDM 4波段DFB激光器芯片	
硅光直流光源	1270/1290/1310/1330nm大功率25/50/70mW 激光器芯片	数据中心100G/200G/400G

资料来源：公司公告，东海证券研究所

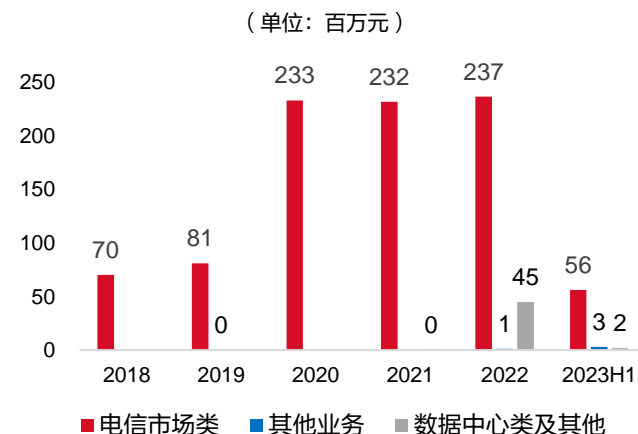
营收、净利润短期承压



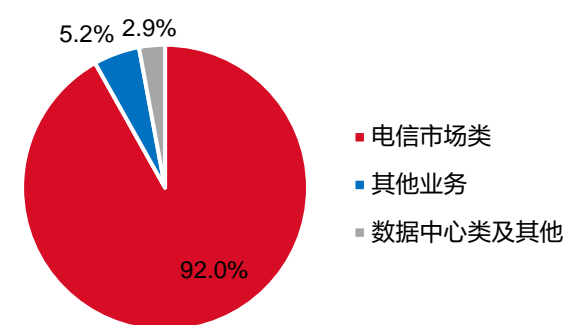
毛利率、净利率短期下滑



电信市场类业务为营收主要贡献力量



电信市场类业务2023年上半年占比超过90%



资料来源：iFind，东海证券研究所

3.8、源杰科技：25G激光器芯片国内率先批量供货，100GEML有望上量

- AI催化数据中心光模块需求暴涨，公司相关高速光芯片有望率先抢占市场，25G激光器芯片实现大批量供货并打破国外垄断。随着云计算、物联网和数字化转型的推进，数据中心和网络设备对高速、高带宽光通信的需求不断增长，数据中心光模块市场也呈现出爆发式增长，尤其是400G/800G以上的需求，同时对功耗和可靠性等也提出了更高的要求。数据中心内的光模块需求将带动相关10G/25G/50G/100G/CW光源等芯片产品的需求，其中高速率光芯片（25G及以上）市场的增长速度将高于中低速率光芯片。国内光芯片市场中，2.5G、10G激光器芯片市场国产化程度较高，但不同波段产品应用场景不同，工艺难度差异大，公司凭借长期技术积累实现激光器光源发散角更小、抗反射光能力更强等差异化特性，为光模块厂商提供全波段、多品类产品，同时提供更低成本的集成方案，实现差异化竞争；25G及更高速率激光器芯片市场国产化率低，公司凭借核心技术及IDM模式，率先攻克技术难关、打破国外垄断，并实现25G激光器芯片系列产品的大批量供货。

源杰科技主要业务进展

业务	进展
激光器芯片	2023年上半年，公司扩建了多条EML芯片生产设备和开发设备，可满足各速率EML芯片的设计、开发、生产，目前已有多个EML产品推向接入网，传输和数据中心AI市场；AI加速了100GEML产品的快速上量，也会加速产品的价格曲线上行和生命周期迭代，公司100GPAM4EML产品处于客户端的测试阶段，有助于突破100GPAM4 EML激光器芯片的海外技术垄断。200GPAM4EML在研发中，200GPAM4EML研发进度符合预期
CW光源产品	早期50mW大功率硅光激光器产品，已经实现销售；目前在产品研发方面，100mW大功率硅光激光器产品几乎可以实现定制化需求，也在逐步向客户送量中
PON产品线	公司研发的50GPONEML激光器和25GPONEML激光器已经和客户进行深化合作，几乎与国际厂商同步
无线产品线方面	市场需求的10G、25G及50G系列DFB激光器产品，公司已经实现送样或出货
磷化铟（InP）集成光芯片方案	下一代数据中心应用400G/800G传输速率方案，传统DFB激光器芯片短期内无法同时满足高带宽性能、高良率的要求，需考虑采用EML激光器芯片以实现单波长100G的高速传输特性。同时，随着应用于数据中心间互联的波分相干技术普及，基于磷化铟（InP）集成技术的光芯片由于具备紧凑小型化、高密集成等特点，可应用于双密度四通道小型可插拔封装（QSFP-DD）等更小型端口光模块，其应用规模将进一步的提升；据C&C的统计，2020年在磷化铟（InP）半导体激光器芯片产品对外销售的国内厂商中，公司收入排名第一，其中10G、25G激光器芯片系列产品的出货量在国内同行业公司中均排名第一

资料来源：公司公告，东海证券研究所

目录

第一部分 大模型带动AI服务器高增长

第二部分 算力芯片与光模块长期受益

第三部分 A股上市公司代表

第四部分 投资建议

第五部分 风险提示



4.1、重点公司业绩一致预期

	营收 (百万元) (YOY)					归母净利润 (百万元) (YOY)					PE			PS	总市值 (亿元)
	2021	2022	2023E	2024E	2025E	2021	2022	2023E	2024E	2025E	2023E	2024E	2025E		
海光信息 (688041.SH)	2310.42 (+126.07%)	5125.27 (+121.83%)	6012.00 (+17.30%)	8426.64 (+40.16%)	11181.73 (+32.69%)	327.11 (+935.65%)	803.54 (+145.65%)	1262.44 (+57.11%)	1667.72 (+32.10%)	2246.93 (+34.73%)	162.11	116.81	86.70	30.81	1852.00
寒武纪 (688256.SH)	721.05 (+57.12%)	729.03 (+1.11%)	709.39 (-2.70%)	1527.68 (+115.35%)	2442.58 (+59.89%)	-824.95 (-89.86%)	-1256.56 (-52.32%)	-835.61 (+33.50%)	-556.69 (+33.38%)	-263.65 (+52.64%)	-87.41	-134.49	-283.96	100.72	714.50
澜起科技 (688008.SH)	2562.02 (+40.49%)	3672.26 (+43.33%)	2285.74 (-37.76%)	4179.33 (+82.84%)	6031.71 (+44.32%)	829.14 (-24.88%)	1299.38 (+56.71%)	450.91 (-65.30%)	1319.12 (+192.55%)	2084.52 (+58.02%)	129.09	45.06	28.51	26.43	604.10
中际旭创 (300308.SZ)	7695.40 (+9.16%)	9641.79 (+25.29%)	10724.79 (+11.23%)	23882.18 (+122.68%)	30710.42 (+28.59%)	876.98 (+1.33%)	1223.99 (+39.57%)	2180.98 (+78.19%)	4201.06 (+92.62%)	5363.97 (+27.68%)	66.37	32.60	25.53	12.35	1325.00
天孚通信 (300308.SZ)	1032.39 (+18.20%)	1196.39 (+15.89%)	1938.99 (+62.07%)	3294.55 (+69.81%)	4647.24 (+41.06%)	306.39 (+9.77%)	402.94 (+31.51%)	729.91 (+81.14%)	1133.85 (+55.34%)	1546.66 (+36.41%)	80.40	48.34	35.44	27.92	541.40
新易盛 (300502.SZ)	2908.38 (+45.57%)	3310.57 (+13.83%)	3107.77 (-6.13%)	5384.94 (+73.27%)	7180.55 (+33.34%)	661.93 (+34.60%)	903.58 (+36.51%)	690.61 (-23.57%)	1278.55 (+87.91%)	1713.28 (+34.00%)	70.85	37.70	28.14	15.20	472.30
光迅科技 (002281.SZ)	6486.30 (+7.28%)	6911.88 (+6.56%)	6520.40 (-5.66%)	7734.40 (+18.62%)	8815.40 (+13.98%)	567.27 (+16.39%)	608.41 (+7.25%)	591.20 (-2.83%)	726.40 (+22.87%)	859.20 (+18.28%)	57.58	46.86	39.62	4.79	312.40
源杰科技 (688498.SH)	232.11 (-0.54%)	282.91 (+21.89%)	144.40 (-48.96%)	304.96 (+111.19%)	432.89 (+41.95%)	95.29 (+20.85%)	100.32 (+5.28%)	19.48 (-80.58%)	96.63 (+396.05%)	143.89 (+48.90%)	323.54	135.68	91.12	86.98	125.60

资料来源: iFind, 各公司公告, 东海证券研究所

(注: 除光迅科技外其他企业2023年营收与归母净利润预期值为其业绩快报中披露的相关数值; 全部企业的总市值为截至2024年3月13日数据; 全部企业的PS为其3月13日总市值与其2023年营收预期的比值)

4.2、半导体行业估值与投资建议

2024/3/13		PE估值				PS估值			PB估值		
指数	代码	板块	PE (TTM)	历史分位数 (5y)	历史分位数 (10y)	PS (TTM)	历史分位数 (5y)	历史分位数 (10y)	PB (MRQ)	历史分位数 (5y)	历史分位数 (10y)
申万电子二级指数	801081.SI	半导体	64.35	46.38%	31.43%	6.51	47.94%	66.72%	4.91	6.43%	14.13%
	801083.SI	电子元器件	32.11	44.32%	26.42%	2.63	57.50%	39.81%	3.24	6.96%	4.15%
	801084.SI	光学光电子	65.15	85.96%	81.66%	1.35	42.17%	21.08%	2.75	27.14%	17.11%
	801085.SI	消费电子	29.73	30.60%	19.62%	1.21	38.71%	20.75%	3.48	11.61%	5.79%
	801086.SI	电子化学品	48.92	47.65%	38.04%	4.70	39.46%	32.58%	3.94	10.87%	11.42%
	801082.SI	其他电子	43.86	51.48%	38.80%	0.96	76.11%	44.49%	4.38	2.59%	2.05%
大盘指数	000001.SH	上证指数	12.97	39.54%	34.22%	1.09	35.71%	28.82%	3.29	11.12%	7.72%
	399001.SZ	深证成指	21.12	5.35%	19.10%	1.47	5.11%	9.98%	2.12	5.19%	12.33%
	399006.SZ	创业板指	28.41	6.10%	3.90%	2.86	5.27%	2.63%	3.76	5.77%	8.09%
	000300.SH	沪深300	11.51	16.47%	22.06%	1.17	25.16%	24.32%	3.49	8.07%	8.30%
行业指数	801080.SI	电子(申万)	44.05	65.65%	46.43%	2.25	72.57%	43.51%	4.02	4.20%	5.65%
	TWSE071.TW	台湾电子指数	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	SOX.GI	费城半导体指数	51.20	99.76%	99.79%	11.94	99.60%	99.66%	9.13	99.05%	99.18%

资料来源：Wind，东海证券研究所（注：该表为截至2024/3/13的周数据）

- **半导体估值：**目前A股沪深300指数的PB处于历史的5年、10年估值分位数的8.07%、8.3%，PE处于5年、10年估值分位数的16.47%、22.06%；半导体指数的PB处于历史的5年、10年估值分位数的6.43%、14.13%，PE处于5年、10年估值分位数的46.38%、31.43%。行业整体的估值水平相对较低，具备长期配置价值。
- **AI算力与光模块估值：**短期内国内算力芯片相关企业还处于追赶海外阶段，国内AI数据中心构建对国产算力芯片有一定的驱动作用，长期高增长有待跟踪。光模块国内市场相对成熟，先进光模块出口海外及国内互联网巨头企业为主。近两年的市场资金追捧，部分AI相关企业的业绩短期兑现难度较大，估值相对较高。
- **投资建议：**建议保持对AI的持续关注，关注市场主题催化行情为主。

目录

第一部分 大模型带动AI服务器高增长

第二部分 算力芯片与光模块长期受益

第三部分 A股上市公司代表

第四部分 投资建议

第五部分 风险提示



5、风险提示

（1）AI需求不及预期风险。目前在大模型的刺激下，全球都在积极布局AI产业链，市场还处于不断投资过程中，收益方式还需要看下游应用场景，消费者的接受意愿。如果下游需求不及预期，对产业的持续投资或将产生影响，从而影响产业链上下游的企业经营业绩。

（2）行业竞争过度风险。随着AI产业的大力投资，不少创业公司也纷纷加入布局AI产业，整个商业模式的行业壁垒相对较高，但全球巨型科技企业纷纷跟随布局，过度的产业竞争或将造成企业经营压力增大，一旦缺少持续性的资金投入，企业或将有经营业绩风险，同时整个产业链或受到冲击。

（3）国际贸易政策的变化风险。AI产业中多个细分市场都需要全球先进科技产品支撑，是全球人类共同努力的成果。然而个别经济体随意更改国际贸易政策，或将导致部分核心产业链断供，对全球其他经济体的AI产业布局产生较大影响。

一、评级说明

	评级	说明
市场指数评级	看多	未来6个月内沪深300指数上升幅度达到或超过20%
	看平	未来6个月内沪深300指数波动幅度在-20%—20%之间
	看空	未来6个月内沪深300指数下跌幅度达到或超过20%
行业指数评级	超配	未来6个月内行业指数相对强于沪深300指数达到或超过10%
	标配	未来6个月内行业指数相对沪深300指数在-10%—10%之间
	低配	未来6个月内行业指数相对弱于沪深300指数达到或超过10%
公司股票评级	买入	未来6个月内股价相对强于沪深300指数达到或超过15%
	增持	未来6个月内股价相对强于沪深300指数在5%—15%之间
	中性	未来6个月内股价相对沪深300指数在-5%—5%之间
	减持	未来6个月内股价相对弱于沪深300指数5%—15%之间
	卖出	未来6个月内股价相对弱于沪深300指数达到或超过15%

二、分析师声明

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，具备专业胜任能力，保证以专业严谨的研究方法和分析逻辑，采用合法合规的数据信息，审慎提出研究结论，独立、客观地出具本报告。

本报告中准确反映了署名分析师的个人研究观点和结论，不受任何第三方的授意或影响，其薪酬的任何组成部分无论是在过去、现在及将来，均与其在本报告中所表述的具体建议或观点无任何直接或间接的关系。

署名分析师本人及直系亲属与本报告中涉及的内容不存在任何利益关系。

三、免责声明

本报告基于本公司研究所及研究人员认为合法合规的公开资料或实地调研的资料，但对这些信息的真实性、准确性和完整性不做任何保证。本报告仅反映研究人员个人出具本报告当时的分析和判断，并不代表东海证券股份有限公司，或任何其附属或联营公司的立场，本公司可能发表其他与本报告所载资料不一致及有不同结论的报告。本报告可能因时间等因素的变化而变化从而导致与事实不完全一致，敬请关注本公司就同一主题所出具的相关后续研究报告及评论文章。在法律允许的情况下，本公司的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，并可能为这些公司正在提供或争取提供多种金融服务。

本报告仅供“东海证券股份有限公司”客户、员工及经本公司许可的机构与个人阅读和参考。在任何情况下，本报告中的信息和意见均不构成对任何机构和个人的投资建议，任何形式的保证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。本公司客户如有任何疑问应当咨询独立财务顾问并独自进行投资判断。

本报告版权归“东海证券股份有限公司”所有，未经本公司书面授权，任何人不得对本报告进行任何形式的翻版、复制、刊登、发表或者引用。

四、资质声明

东海证券股份有限公司是经中国证监会核准的合法证券经营机构，已经具备证券投资咨询业务资格。我们欢迎社会监督并提醒广大投资者，参与证券相关活动应当审慎选择具有相当资质的证券经营机构，注意防范非法证券活动。

上海 东海证券研究所

地址：上海市浦东新区东方路1928号 东海证券大厦

网址：[Http://www.longone.com.cn](http://www.longone.com.cn)

座机：（8621）20333275

手机：18221959689

传真：（8621）50585608

邮编：200215

北京 东海证券研究所

地址：北京市西三环北路87号国际财经中心D座15F

网址：[Http://www.longone.com.cn](http://www.longone.com.cn)

座机：（8610）59707105

手机：18221959689

传真：（8610）59707100

邮编：100089

T H A N K
务实

创新

Y O U
规范

协同



东海证券



东海研究