科技专题研究

2024年01月07日

中航证券有限公司
AVIC SECURITIES CO., LTD.

AI智算时代已至，算力芯片加速升级

行业评级：增持

相关报告：
《AI大模型开启新一轮大国竞争，半导体战略地位凸显》
《半导体行业深度：后摩尔时代新星，Chiplet与先进封装风云际会》

分析师：刘牧野
证券执业证书号：S0640522040001

核心观点

- **AI正处史上最长繁荣大周期，生态加速收敛：**在进入21世纪以来，在大数据和大算力的支持下，归纳统计方法逐渐占据了人工智能领域的主导地位，深度学习的浪潮席卷人工智能，人工智能迎来史上最长的第三次繁荣期。**智算中心的发展基于最新人工智能理论和领先的人工智能计算架构，当前算法模型的发展趋势以AI大模型为代表，算力技术与算法模型是其中的核心关键，算力技术以AI芯片、AI服务器、AI集群为载体。**
- **GPU主宰算力芯片，AI信创驱动国产算力发展：**得益于硬件支持与软件编程、设计方面的优势，CPU+GPU成为了目前应用最广泛的平台。AI 分布式计算的市场主要由算力芯片 (55-75%)、内存 (10-20%) 和互联设备（10-20%）三部分组成。美国已限制对华销售最先进、使用最广泛的AI训练GPU—英伟达 A100以及H100，国产算力芯片距离英伟达最新产品存在较大差距，但对信息颗粒度要求较低的推理运算能实现部分替代。
- **提升算力内存带宽，HBM供不应求：**由于ChatGPT的爆火，GPU需求明显，英伟达也加大对三星和SK海力士HBM3的订单。2023年10月，SK海力士表示，已经在2023年出售了明年HBM3和HBM3E的所有产量。据Omdia预测，到2025年，HBM市场的总收入将达到25亿美元。
- **集成算力与存力，先进封装产能紧缺：**CoWoS封装技术是目前集成HBM与CPU/GPU处理器的主流方案。台积电主导全球CoWoS封装市场。据IDC预测，全球CoWoS供需缺口约20%，2024年台积电的CoWos封装产能将较2023年提升一倍，2.5D/3D先进封装市场规模在2023-2028年将以22%的CAGR高速增长。
- **AI算力对高效电源提出新需求，背面供电技术蓄势待发：**越来越高度化的集成会造成针对加速芯片的电源解决方案越来越复杂，方案需要不同电压、不同路的多路输入，这种情况下电压轨会越来越多。台积电、三星、英特尔等芯片大厂都在积极布局背面供电网络技术，为日益复杂的芯片提供高效供电方案，其中英特尔较为领先。
- **建议关注：GPU：**海光信息、寒武纪，和未上市的地平线、黑芝麻、摩尔线程；**HBM：**香农芯创、雅克科技；**先进封装：**兴森科技、华海诚科、艾森股份；**电源芯片：**希荻微。
- **风险提示：**AI算法、模型存较高不确定性，AI技术发展不及预期；ChatGPT用户付费意愿弱，客户需求不及预期；针对AI的监管政策收紧

目 录

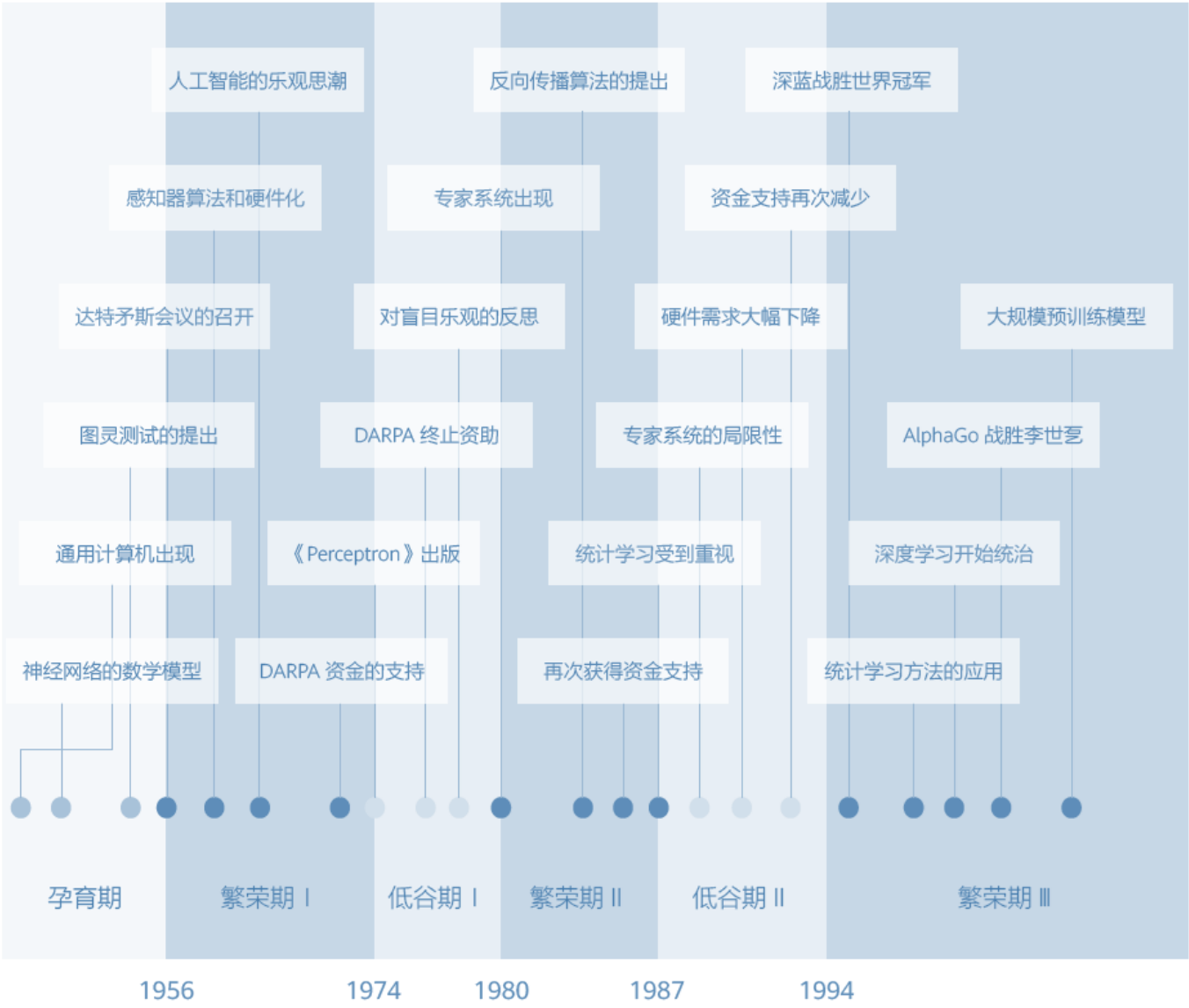
AI正处史上最长繁荣大周期

- 人工智能从1956 年被正式提出以来，经历了数十年的发展历程。人工智能诞生初期，其研究主要分为三个流派，即逻辑演绎、归纳统计和类脑计算。

- 人工智能研究的三大流派各有优劣势。类脑计算流派的目标最为宏远，但在未得到生命科学的支撑之前，难以取得实际应用。归纳演绎流派的思考方式与人类相似，具有较强的可解释性。由于对数据和算力的依赖较少，归纳演绎流派成为人工智能前两次繁荣的主角。随着学界对人工智能困难程度的理解逐渐加深，数理逻辑方法的局限性被不断放大，并最终在第三次繁荣期中，逐渐让位于统计学习的“暴力美学”。

- **在进入21世纪以来，在大数据和大算力的支持下，归纳统计方法逐渐占据了人工智能领域的主导地位，深度学习的浪潮席卷人工智能，人工智能迎来史上最长的第三次繁荣期，至今仍未有结束的趋势。**

图：人工智能发展史



资料来源：华为，中航证券研究所

AI生态加速收敛

- 大模型技术逐步收敛，生态走向聚合，模型更收敛 、 框架更归一 。
- 为了开发更高性能的 AI大模型需要更强的算力平台，算力底座技术门槛将提高，未来训练核心拼集群系统能力 。

图：AI技术逐步收敛，生态走向聚合

	模型更收敛	框架更归一
大模型	基础大模型10~20个	Pytorch框架占95%+
小模型	数万小模型 字节有上万个自研模型	PyTorch、TensorFlow， Caffe，飞桨等

图：算力底座技术门槛提高

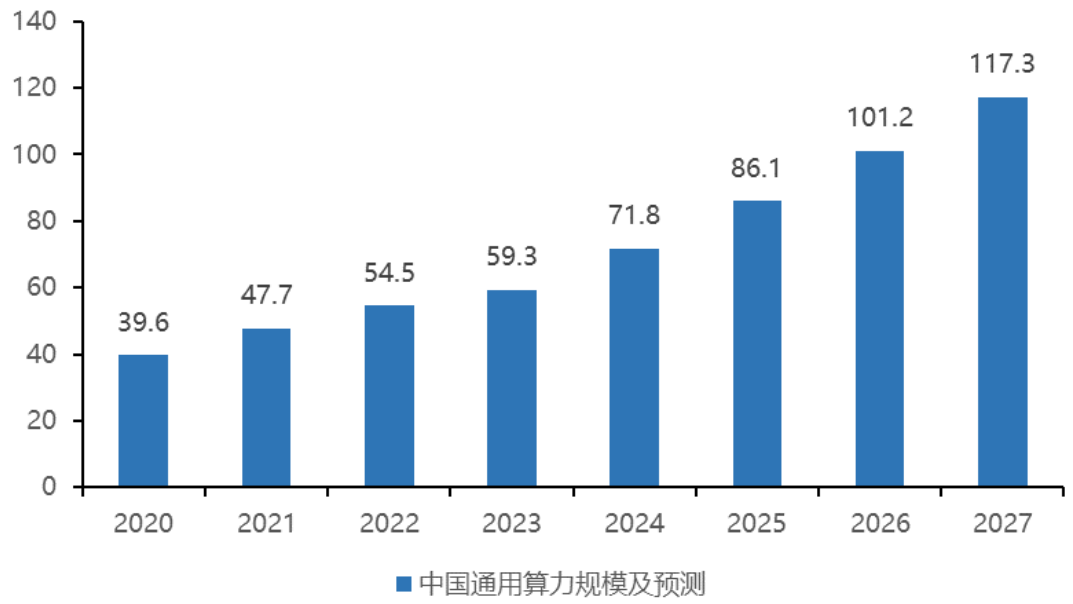
	万级参数时代		亿级参数时代		万亿参数时代
计算需求	百TF级平台 1张GPU卡	X10倍 →	PF级平台 单服务器，8卡	X1000倍 →	EF级平台 AI集群，~万卡
网络需求	无互联	N/A →	节点内卡间互联	X100倍 →	超节点+网络互联
存储需求	GB级存取 -服务器硬盘	X100倍 →	TB级存取 -服务器硬盘	X1000倍 →	PB级存取 -高并发多级存储

资料来源：华为，中航证券研究所

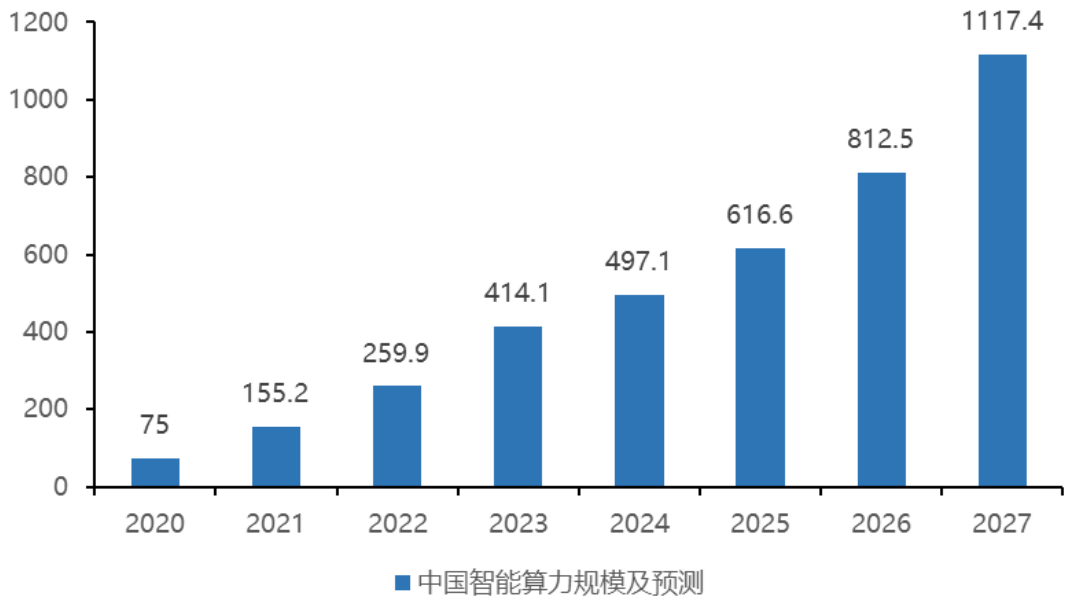
智能算力规模将快速增长

- 复杂的模型和大规模的训练需要大规模的高算力支持，这不仅需要消耗大量计算资源，而且对算力的速度、精度、性能也提出更高要求。
- **市场对于更高性能的智能算力需求将显著提升，智能算力增长速率约通用算力的两倍。**据IDC和浪潮信息测算，2022年中国通用算力规模达 54.5 EFLOPS，预计到2027年通用算力规模将达到117.3 EFLOPS。2022年中国智能算力规模达259.9EFLOPS，预计到2027年将达到 1117.4 EFLOPS。**2022 -2027年期间，中国智能算力规模年复合增长率达33.9%，同期通用算力规模年复合增长率为16.6%。**

图：中国通用算力规模及预测（EFLOPS，基于FP64计算）



图：中国智能算力规模及预测（EFLOPS，基于FP16计算）

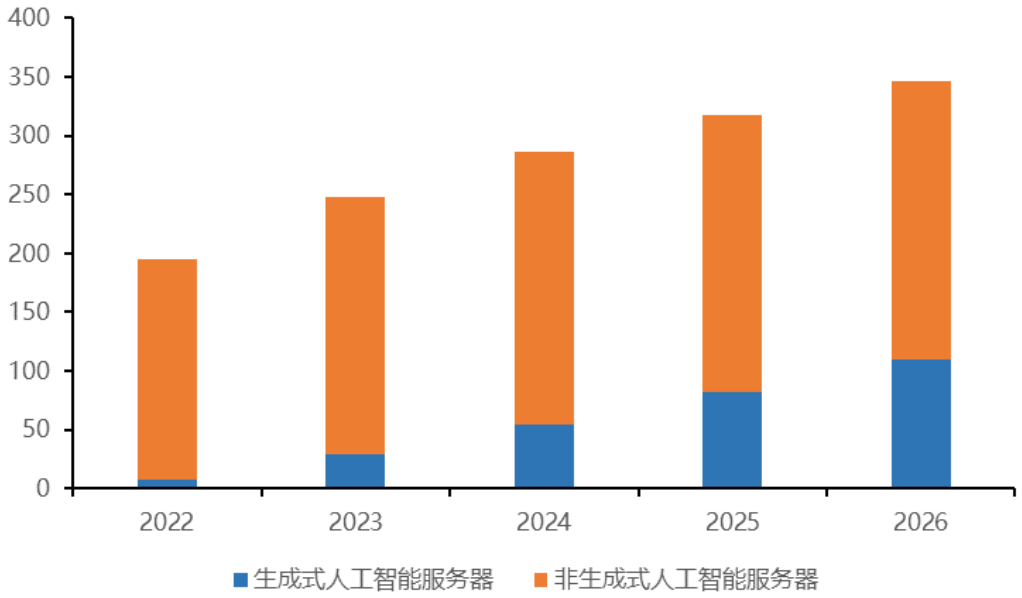


资料来源：浪潮信息，IDC，中航证券研究所

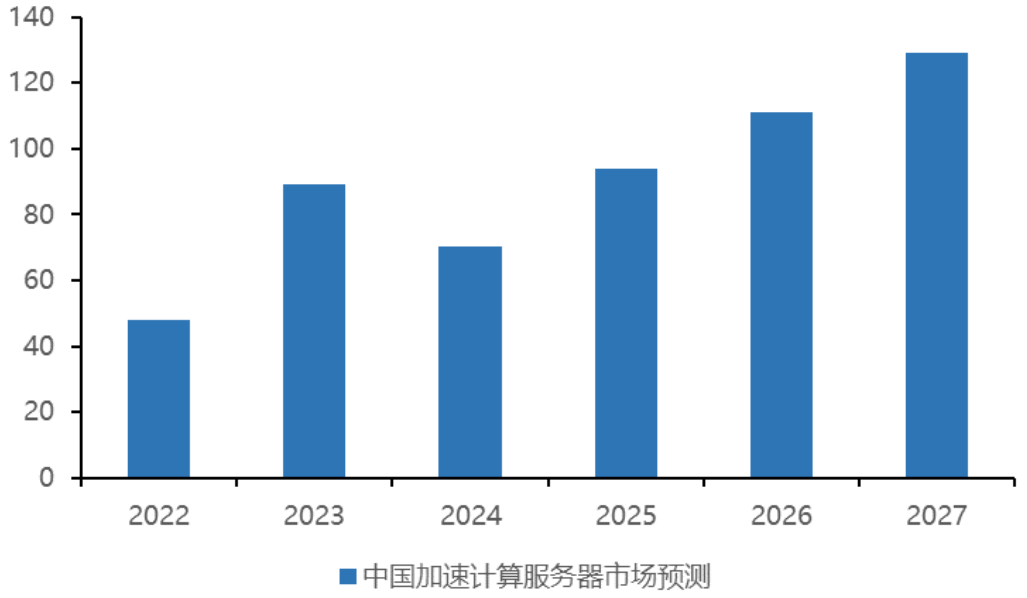
AI服务器需求旺盛

- 从感知智能到生成式智能，人工智能越来越需要依赖“强算法、高算力、大数据”的支持。模型的大小、训练所需的参数量等因素将直接影响智能涌现的质量，人工智能模型需要的准确性越高，训练该模型所需的计算力就越高。IDC预计，全球人工智能硬件市场（服务器）规模将从2022年的195亿美元增长到2026年的347亿美元，五年年复合增长率达17.3%；在中国，预计2023年中国人工智能服务器市场规模将达到91亿美元，同比增长82.5%，2027年将达到134亿美元，五年年复合增长率达21.8%。

图：全球AI计算服务器市场规模预测（亿美元）



图：中国AI计算服务器市场预测（亿美元）



资料来源：浪潮信息，IDC，中航证券研究所

全国推进算力建设，加大算力投资

- **在适度超前的指导思想下，国家正加大对人工智能算力基础设施的投资。**算力基础设施建设成为一个重要环节，被纳入国家新基建范畴。据IDC统计，截至2023年8月，全国已有超过30个城市建设智算中心，总建设规模超过200亿。

图：国内算力规划

2023年4月
《上海市推进算力资源统一调度指导意见》

到2023年，可调度智能算力达到1000 PFLOPS（FP16）以上；到2025年，本市数据中心算力超过18000 PFLOPS（FP32）

2023年10月
《算力基础设施高质量发展行动计划》

到 2025 年，计算力方面，全国算力规模超过300 EFLOPS，智能算力占比达到 35%，东西部算力平衡协调发展。

2023年12月
《深圳市算力基础设施高质量发展行动计划》

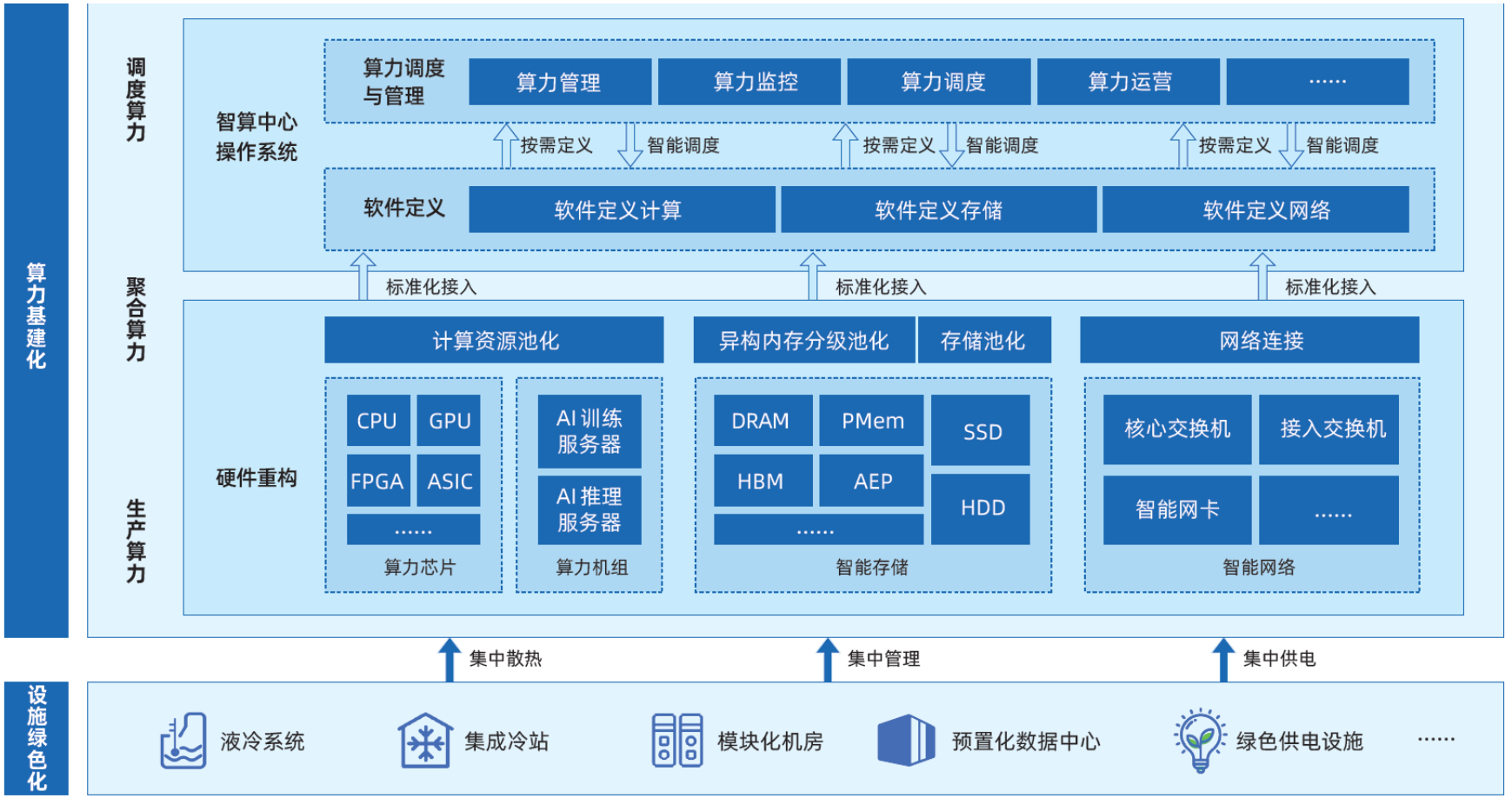
到2025年，通用算力达到14EFLOPS（FP32），智能算力达到25EFLOPS（FP16），超算算力达到2EFLOPS（FP64）。

资料来源：浪潮信息，IDC，政府官网，中航证券研究所

算力、存储、网络构建智算中心基础

- 智算中心的发展基于最新人工智能理论和领先的人工智能计算架构，算力技术与算法模型是其中的核心关键，算力技术以AI芯片、AI服务器、AI集群为载体，而当前算法模型的发展趋势以AI大模型为代表。

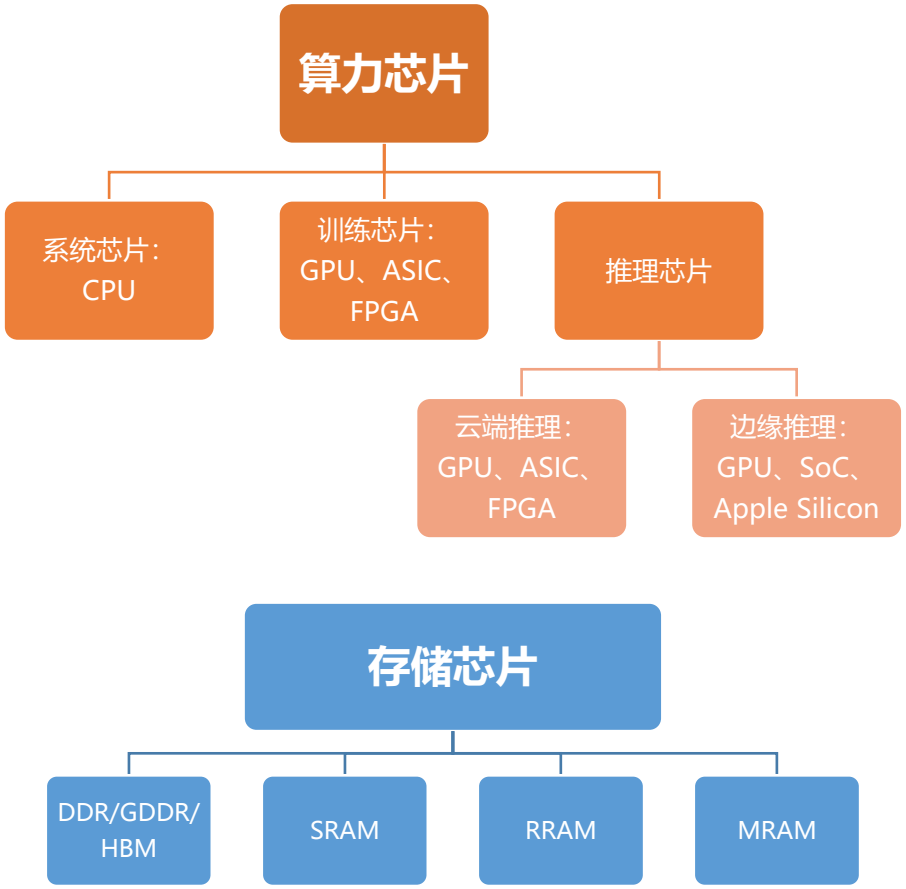
图：智算中心算力基础架构



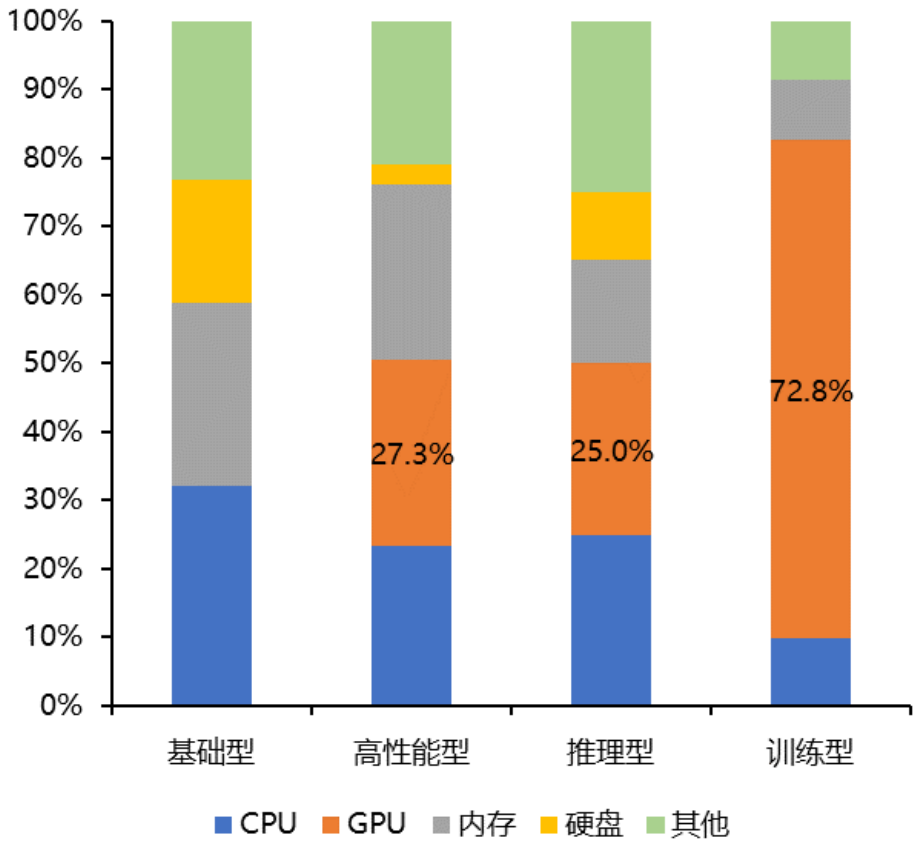
资料来源：国家信息中心《智能计算中心创新发展指南》，中航证券研究所

算力芯片主导AI计算市场

- AI 分布式计算的市场主要由算力芯片 (55-75%)、内存 (10-20%) 和互联设备（10-20%）三部分组成。美国已限制对华销售最先进、使用最广泛的AI训练GPU—英伟达 A100以及H100，国产算力芯片距离英伟达最新产品存在较大差距，但对信息颗粒度要求较低的推理运算能实现部分替代。
- **GPU占AI服务器成本最高，国产替代重要性凸显。**我们认为，AI训练芯片受限进一步强调了高制程芯片设计、代工的国产替代紧迫性。



图：各类型服务器成本结构占比情况

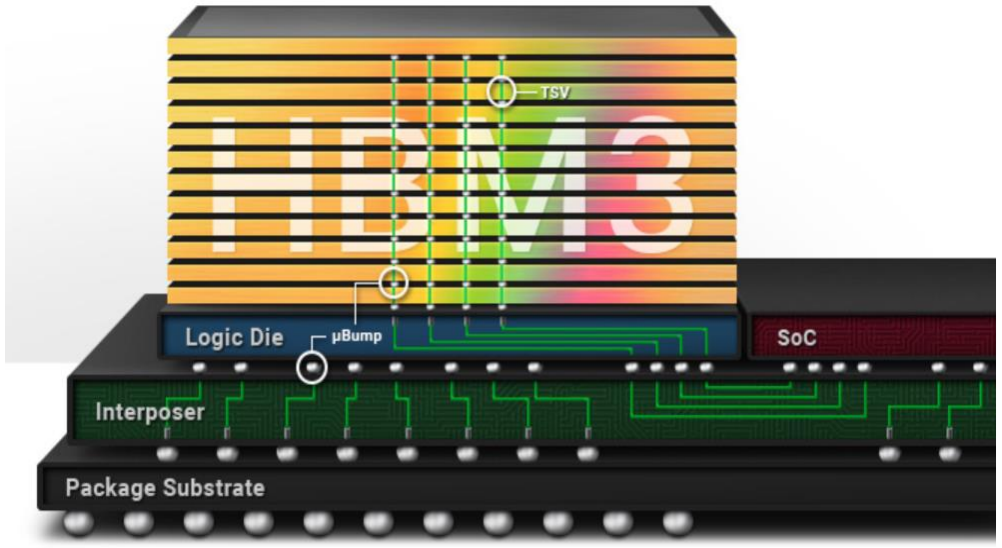


资料来源：拾象科技，华经产业研究，中航证券研究所

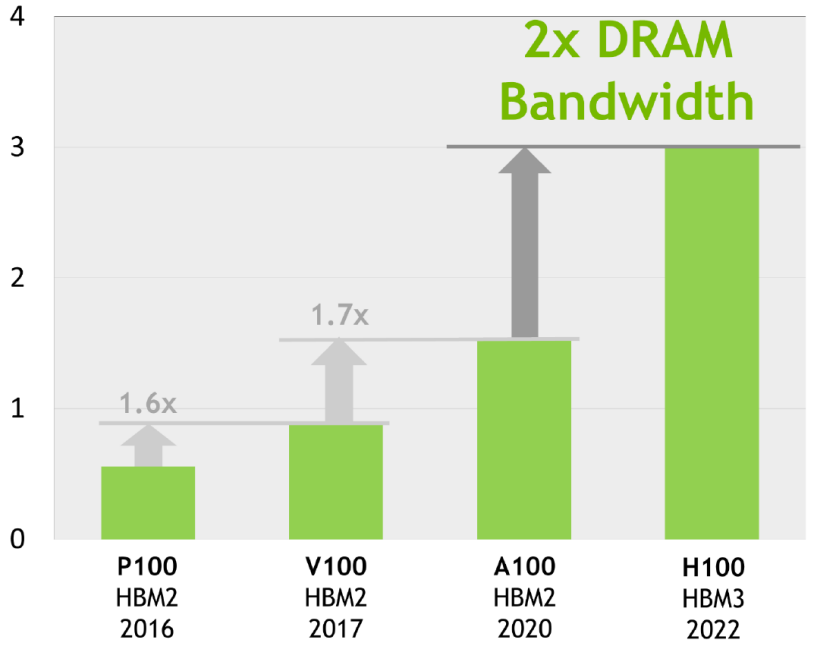
提升算力内存带宽，HBM供不应求

- 以ChatGPT为代表的生成类模型需要在海量数据中训练，对存储容量和带宽提出新要求，HBM（High Bandwidth Memory，高带宽存储器）成为减小内存墙的优选项。HBM将多个DDR芯片堆叠并与GPU封装在一起，是一种基于3D堆叠工艺的高附加值DRAM产品。通过增加带宽，扩展内存容量，让更大模型、更多参数留在离计算核心区更近的地方，从而减少内存和存储解决方案带来的延迟。据Omdia预测，到2025年，HBM市场的总收入将达到25亿美元。
- 由于ChatGPT的爆火，GPU需求明显，英伟达也加大对三星和SK海力士HBM3的订单。2023年10月，SK海力士表示，已经在2023年出售了明年HBM3和HBM3E的所有产量。

图：HBM3产品结构



图：英伟达使用的HBM带宽不断升级（TB/s）

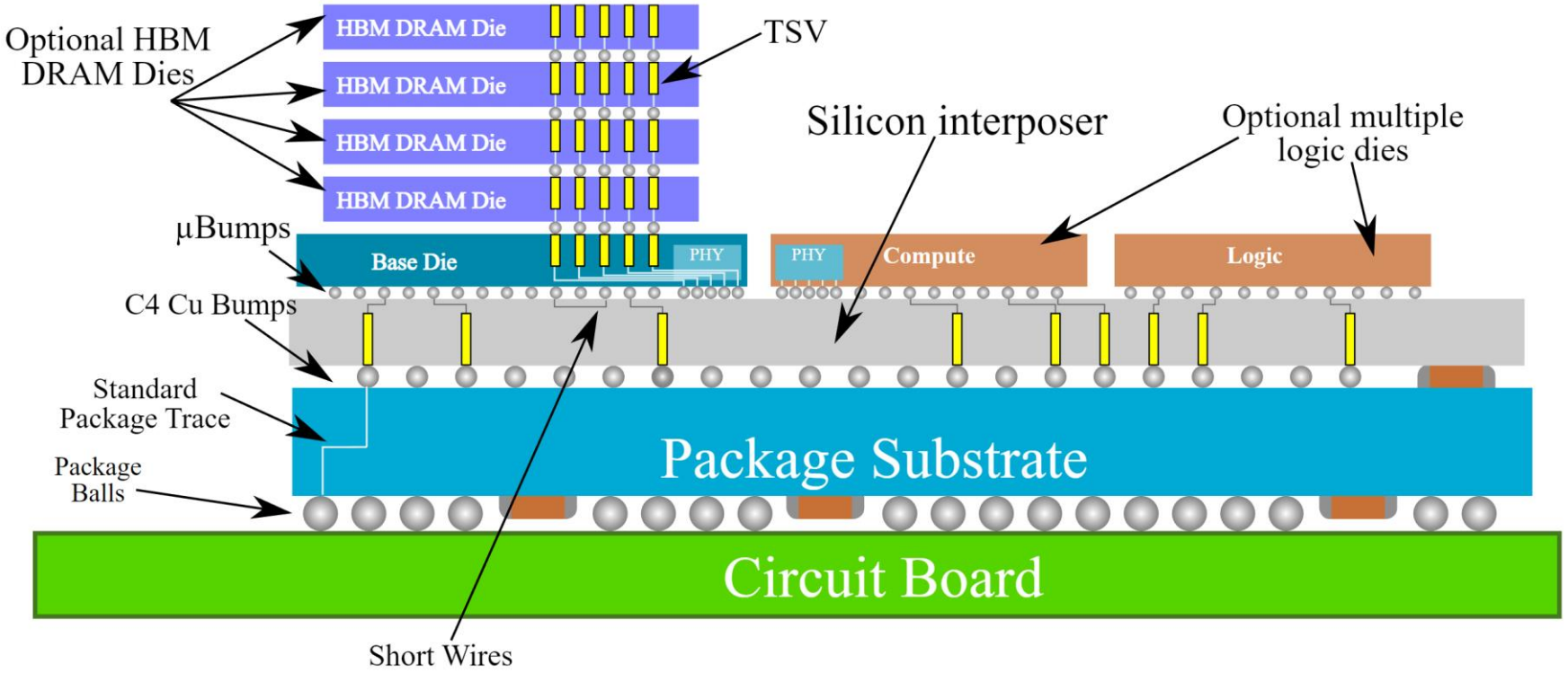


资料来源：SK海力士，英伟达，中航证券研究所

集成算力与存力，先进封装产能紧缺

- **CoWoS封装技术是目前集成HBM与CPU/GPU处理器的主流方案。**在算力芯片性能暴增的时代下，先进封装产业链逐渐的进入高速发展时期。
- **台积电封装产能紧缺。**台积电主导全球CoWoS封装市场，且正在扩大产能，以满足客户，尤其是AI芯片领域的需求。英伟达等大客户增加了对CoWoS封装的订单量，AMD、亚马逊等其他大厂也出现了紧急订单。据IDC预测，全球CoWoS供需缺口约20%，2024年台积电的CoWos封装产能将较2023年提升一倍，2.5D/3D先进封装市场规模在2023-2028年将以22%的CAGR高速增长。

图：台积电CoWoS封装

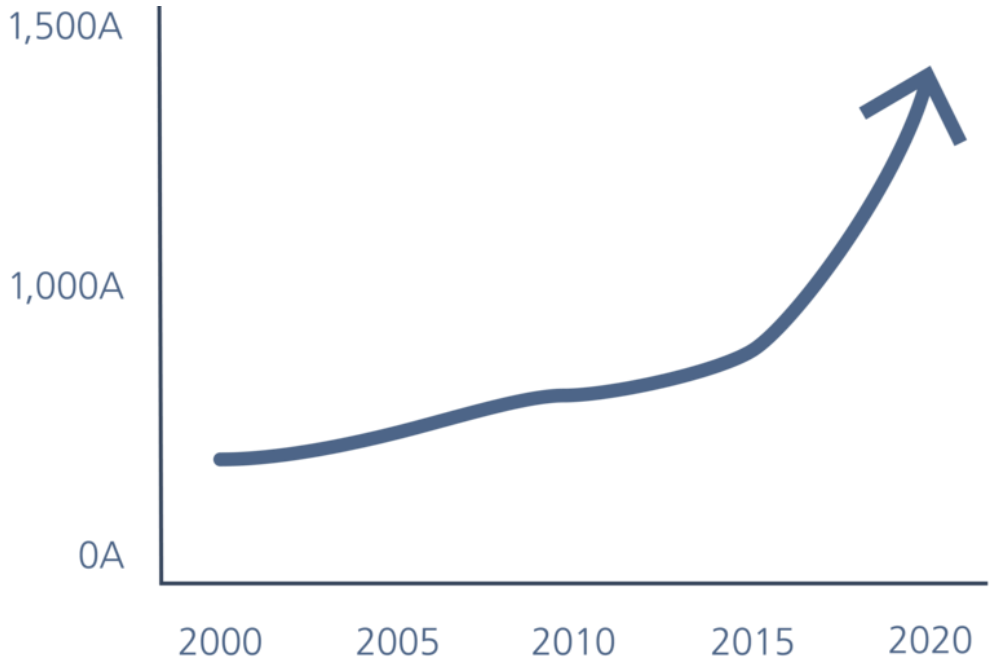


资料来源：semiconductor engineering，中航证券研究所

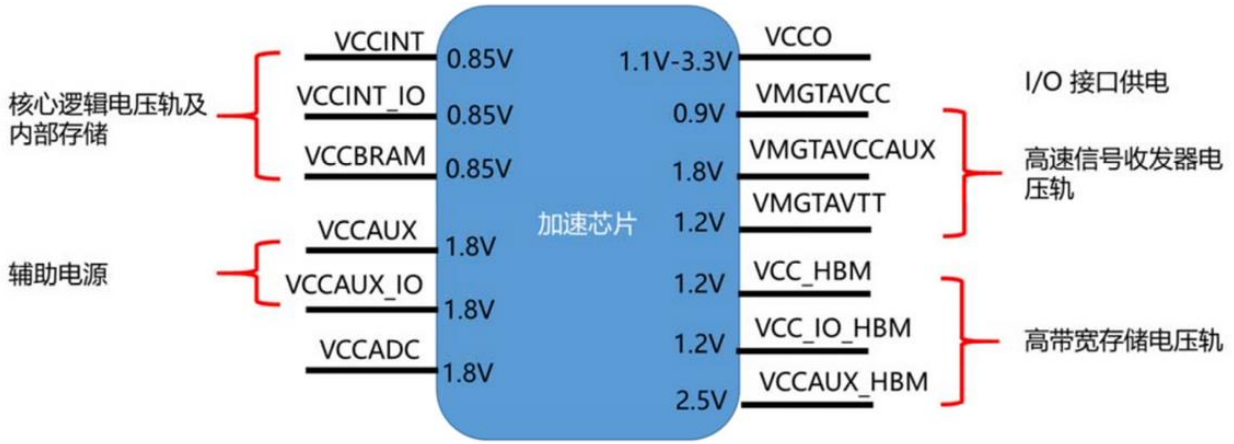
AI算力对高效电源提出新需求

- **AI算力功耗增长。**当代 GPU 有数百亿颗晶体管，更好的处理性能是以指数级增长的电源需求为代价的，因此人工智能和机器学习等应用的高性能处理器需要不断增加功率。据vicorpower，目前的趋势是处理器的功耗每两年翻一番，2000A 的峰值电流现在已经很普遍。
- **AI芯片供电架构愈发复杂。**越来越高度化的集成会造成针对加速芯片的电源解决方案越来越复杂，方案需要不同电压、不同路的多路输入，这种情况下电压轨会越来越多。

图：GPU/CPU 峰值电流提升



图：AI芯片供电架构日趋复杂



核电压降低 **大动态** **多路输入** **电流随应用变化**

资料来源：vicorpower，MPS，中航证券研究所

算力产业链面临国产化机会和挑战

■ AI算力芯片处于AI计算的最上游，GPU、HBM、先进封装等环节需求高增，甚至已出现供不应求的现象。目前算力芯片产业链由海外公司主导，在美国制裁中国科技发展，限制半导体技术输入中国的背景下，AI算力芯片在各环节均存在需求扩张叠加国产替代的双重增长动力。

GPU	HBM	先进封装	电源管理芯片
➢国外 • 英伟达、AMD ➢国内 • 海光信息、寒武纪、龙芯中科、摩尔线程、燧原科技	➢国外 • 海力士、三星、美光 ➢国内 • 香农芯创（经销商）、雅克科技（原材料）	➢国外 • 英特尔、三星 ➢国内 • 封装厂：台积电、盛合晶微、长电科技、通富微电、甬矽电子 • 封装材料：南电、欣兴、兴森科技、联瑞新材、生益科技	➢国外 • MPS、德州仪器、ADI ➢国内 • 希荻微、杰华特、晶丰明源

资料来源：中航证券研究所

目 录

AI芯片架构众多

- AI芯片根据其技术架构，可分为GPU、FPGA、ASIC及类脑芯片，同时CPU可执行通用AI计算。相较于传统的中央处理器（CPU），GPU具有并行计算、高效能和高并发等优势，因此在人工智能、机器学习、数据挖掘等领域得到广泛应用。
- AI芯片根据其在网络中的位置可以分为云端AI芯片、边缘及终端AI芯片；根据其在实践中的目标，可分为训练芯片和推理芯片。
- 云端主要部署训练芯片和推理芯片，承担训练和推理任务，具体指智能数据分析、模型训练任务和部分对传输带宽要求比高的推理任务；边缘和终端主要部署推理芯片，承担推理任务，需要独立完成数据收集、环境感知、人机交互及部分推理决策控制任务。

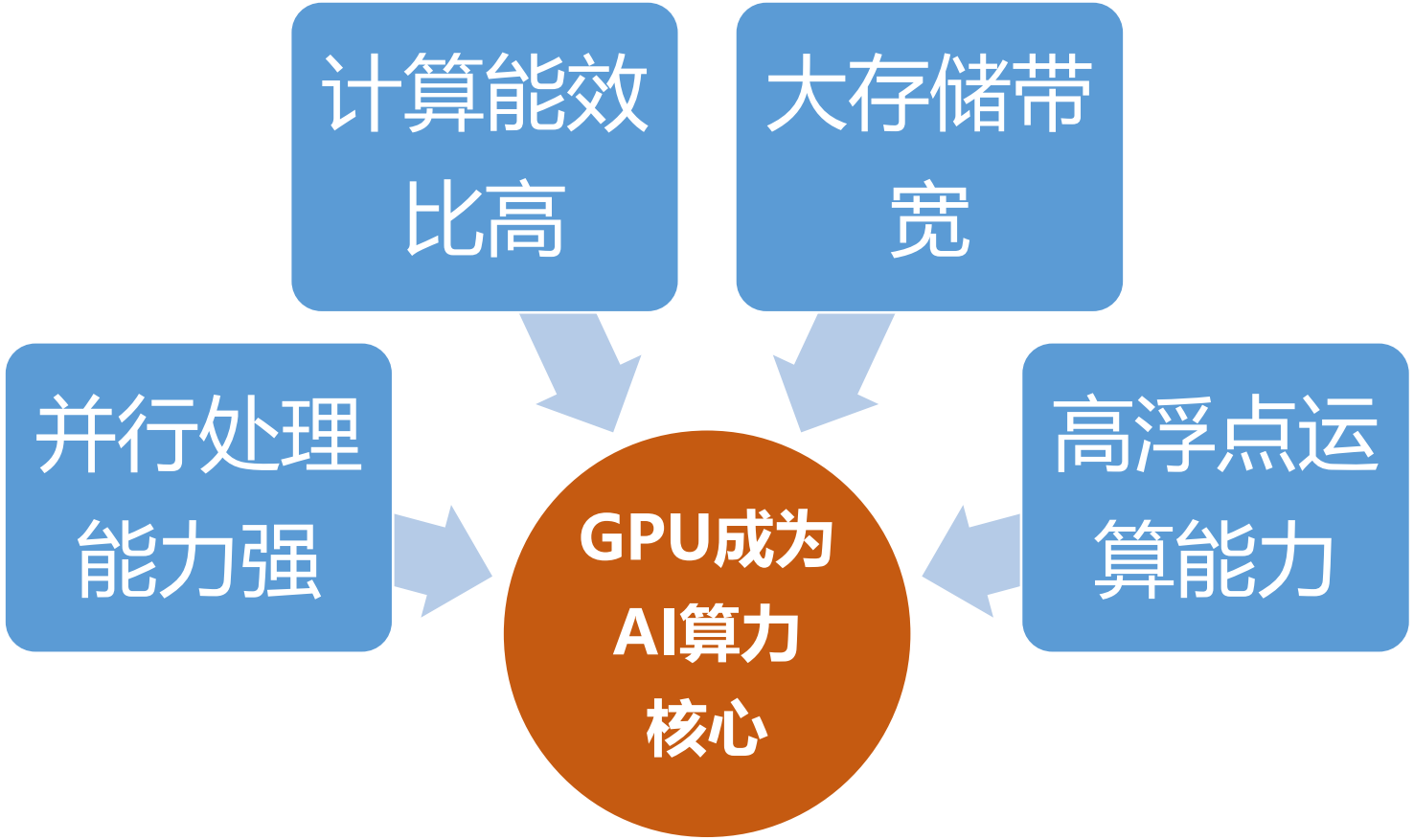
图：AI芯片分类

	芯片架构	芯片特点	代表公司	专用性 (L1到L5依次增强)
发展方向一：从通用到专用	CPU	CPU的通用架构设计使运行效率受限。当前CPU虽然在机器学习领域的计算大大减少,但是不会被完全取代。	英特尔	L1
	GPU	**目前商用最广泛的AI芯片**, 可以执行深度学习和神经网络任务。GPU主要从事大规模并行计算,比CPU运行速度快,并且比其他专用AI处理器芯片价格低。	英伟达、AMD	L2
	DSP	仅作为处理器IP核使用。目前基于DSP的设计有一定的局限性,一般都是针对图像和计算机视觉的处理器IP核芯片,速度较快,成本不高。	新思科技、Cadence	L3
	FPGA	FPGA具有三大优点:单位能耗比低、硬件配置灵活、架构可调整。但是,FPGA的 使用有一定门槛,要求使用者具备硬件知识。	赛灵思、微软	L4
	TPU /ASIC	当前为谷歌公司专用,还不是市场化产品。ASIC芯片不能像FPGA很快改变架构,适应变化,对企业而言成本较昂贵。	谷歌	L5
发展方向二：颠覆经典冯氏架构，采用人脑神经元的结构来提升计算能力	TrueNorth	模仿人脑神经元和神经突触的结构,功耗非常低。有可能实现人工智能领域的通用化路径,但从短期来看,离大规模商业生产还有很远的距离。	IBM	

资料来源：亿欧智库，松鼠厂，中航证券研究所

GPU与深度学习技术相契合

- GPU设计之初用于对图形进行渲染，需要并行处理海量数据，涉及大量矩阵运算。深度学习依赖于数学和统计学计算，所以图形渲染与深度学习有着相似之处，这两种场景都需要处理每秒大量的矩阵乘法运算。GPU拥有数千个内核的处理器，能够并行执行数百万个数学运算。因此GPU完美地与深度学习技术相契合。使用GPU做辅助计算，能够更快地提高AI的性能。



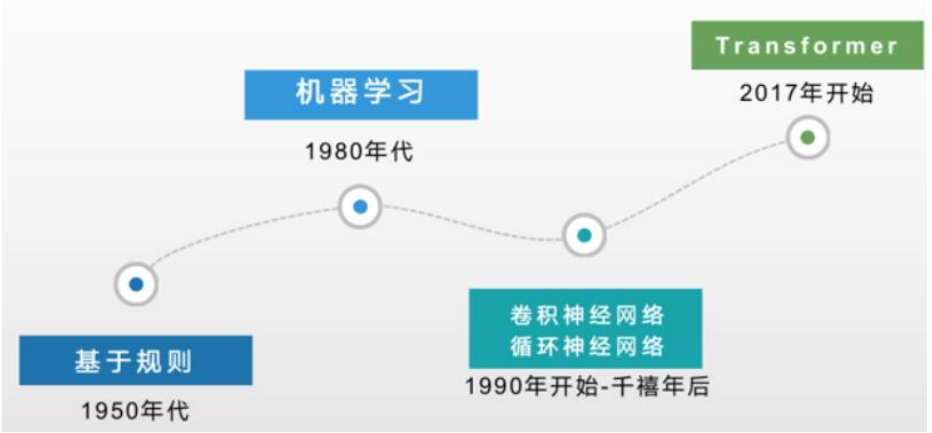
资料来源：知乎，沐曦集成电路，中航证券研究所

AI算力路线向GPU收敛

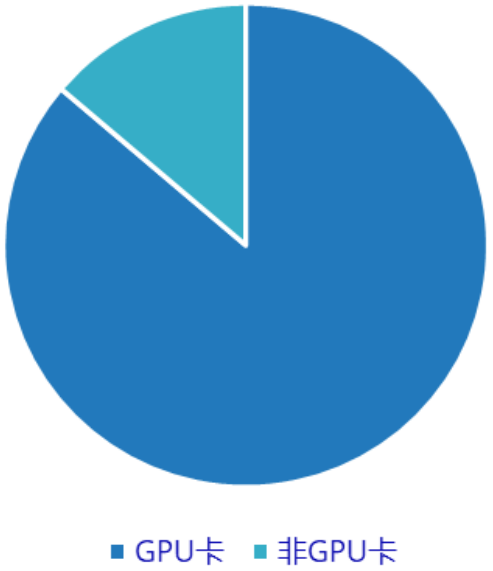
- **大模型的基础架构向 Transformer 结构收敛。** Transformer 结构在图 、 文 、 音多领域表现优异，大量基于 Transformer 结构的大模型涌现 。 Transformer 模型预训练由多个堆叠的自注意力层和前馈神经网络层组成，这种设计使得它在构造大型深度神经网络时具有巨大优势。BERT 和GPT 是两种最知名的基于Transformers 的自然语言处理模型。

- **大模型的发展，使得算法技术路线逐渐统一于Transformer模型，算力芯片技术路线也向适合并行计算的GPU收敛。** Transformer 架构具有并行计算的能力，可以同时处理输入序列的不同部分。在使用分布式计算和GPU并行计算的情况下，Transformer可以更快地训练和推理大型深度神经网络。大模型需要大算力和大互联，对底层GPU支撑规模提出了空前的要求，达到万卡级别。因此，出于对大模型的训练需求，市场选择了GPU作为主流的算力芯片。据IDC，在中国人工智能芯片市场，GPU占有超过80%的市场份额。

图：人工智能进入transformer时代



图：中国人工智能芯片市场份额

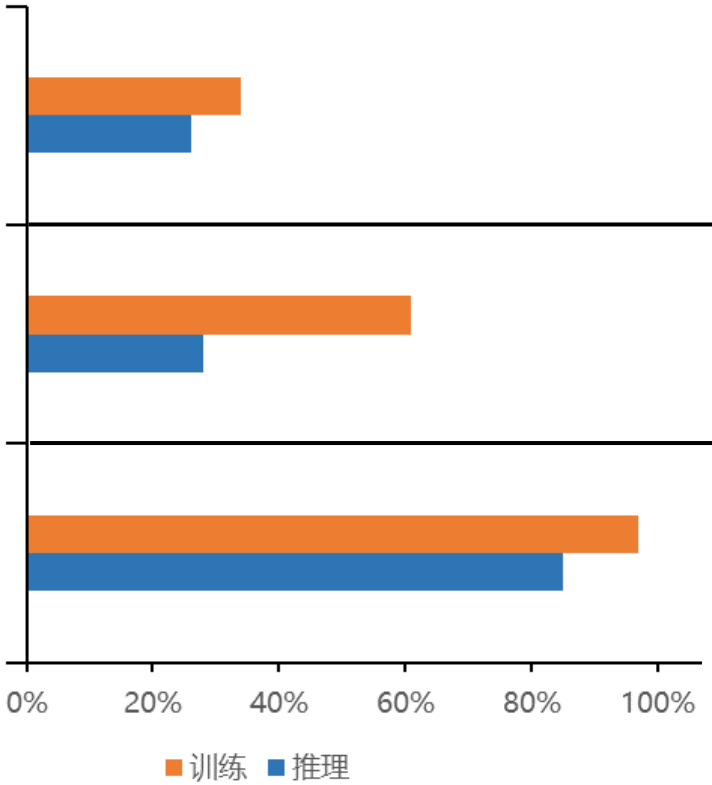


资料来源：IDC，中国移动通信研究院，中航证券研究所

CPU+GPU是人工智能异构计算的主要组合形式

- **异构计算仍然是芯片发展趋势之一。**异构计算通过在单一系统中利用不同类型的处理器（如CPU、GPU、ASIC、FPGA、NPU等）协同工作，执行特定任务，以优化性能和效率，更高效地利用不同类型的计算资源，满足不同的计算需求。
- 得益于硬件支持与软件编程、设计方面的优势，CPU+GPU成为了目前应用最广泛的平台。截至2023年10月，中国市场普遍认为“CPU+GPU”的异构方式是人工智能异构计算的主要组合形式。

图：人工智能训练和推理工作负载选用的计算架构



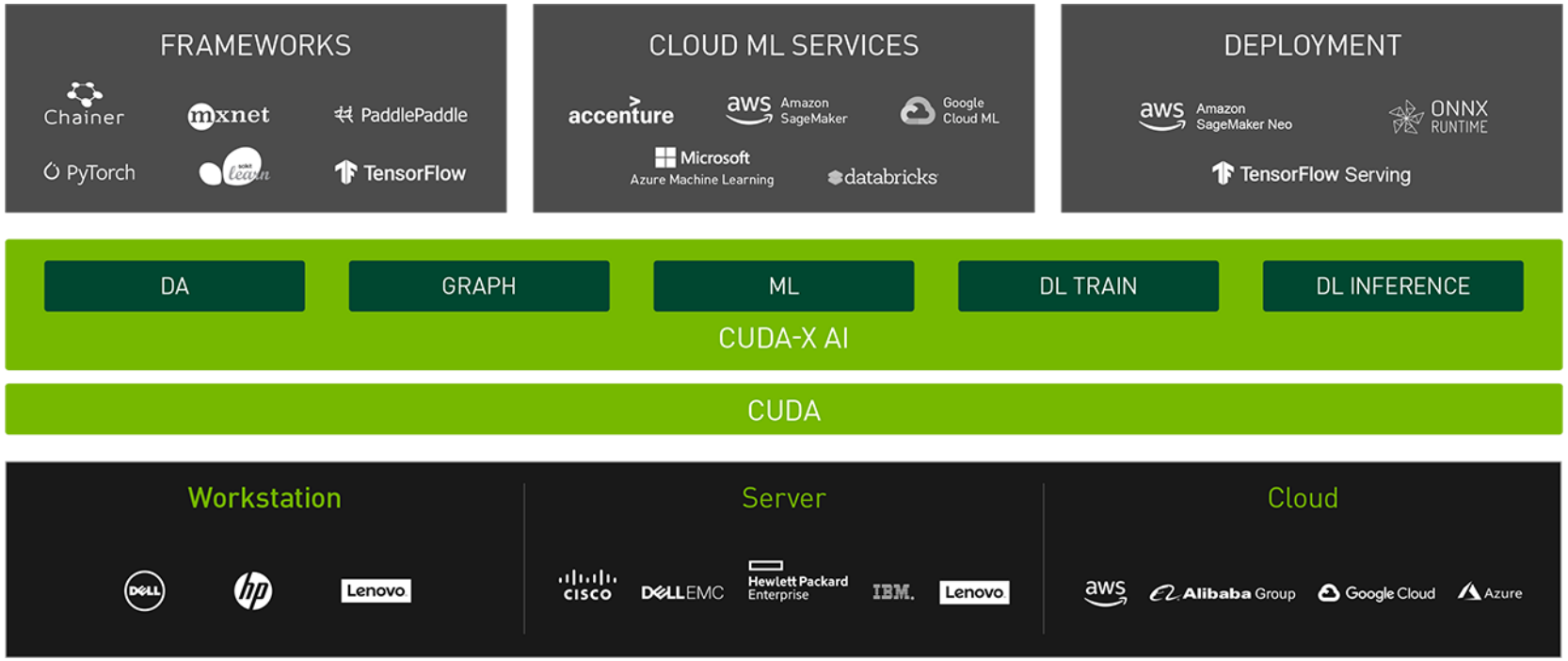
	优点	缺点	实例
CPU+FPGA	FPGA灵活性高、无需取指令、译码，执行效率高	总体性价比和效率不占优势。FPGA的大规模开发难度偏高，从业人员相对较少，生态环境不如GPU	微软使用FPGA为Bing搜索智能化进行加速
CPU+AISC	ASIC体积小、功耗低、计算性能高、计算效率高、芯片出货量越大成本越低	算法固定，一旦算法变化就无法使用。目前人工智能算法远没有到算法平稳期，ASIC专用芯片如何做到适应各种算法是个最大的问题	寒武纪的NPU、地平线的BPU、Google的TPU
CPU+GPU	GPU具有更好的生态环境，可以直接部署在服务器中，方便快速	缺乏灵活性	英伟达、AMD的GPU

资料来源：IDC，浪潮信息，中航证券研究所

坚守CUDA开发生态，奠定英伟达AI王位

- CUDA（Compute Unified Device Architecture，统一计算架构）是由英伟达于2006年所推出的一种软硬件集成技术，是该公司对于GPGPU的正式名称。透过这个技术，GPU的计算能力由图像处理拓展到更广泛的计算领域。CUDA亦是首次可以利用GPU作为C语言编译器的开发环境。
- **现在主流的深度学习框架基本基于CUDA进行GPU并行加速。** CUDA推出之前，给GPU编程需要用机器码深入到显卡内核才能完成任务，非常困难。英伟达推出CUDA后，把复杂的显卡编程包装成了一个简单的接口，降低了开发难度，为广大开发人员提供了简单易学的开发平台。在英伟达十几年坚持不懈推广下，CUDA的高性能运算研究成果陆续在众多知名期刊发表并获得认可，并广泛应用于深度学习、自动驾驶以及其他 AI 领域。

图：CUDA-X AI 软件加速库帮助现代 AI 应用程序加速运行

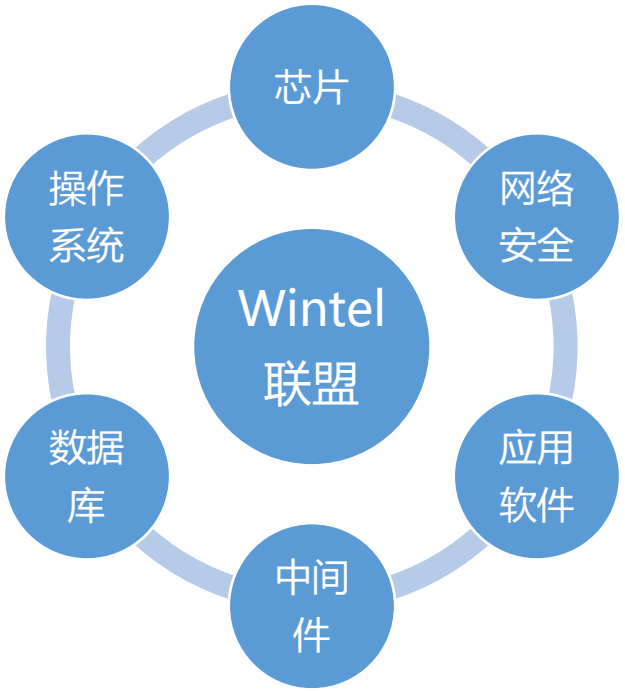


资料来源：英伟达，中航证券研究所

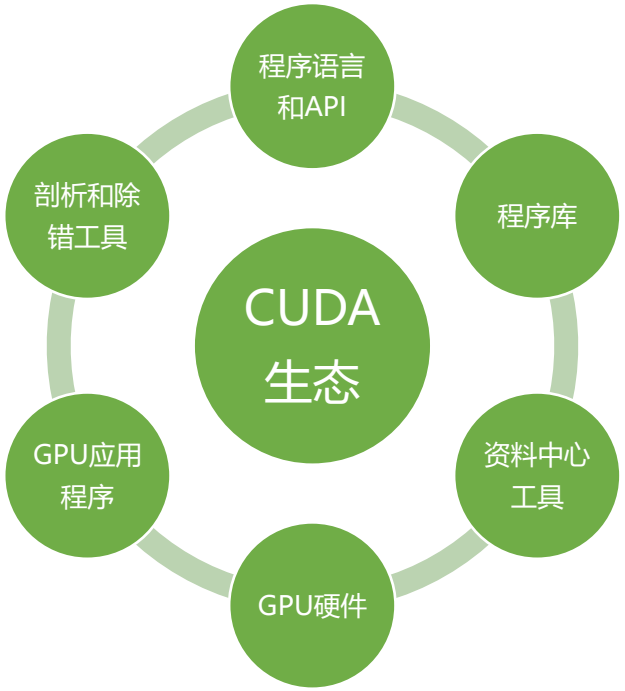
坚守CUDA开发生态，奠定英伟达AI王位

- CUDA处在软硬件结合的关键位置，是软件生态的基石。英伟达的竞争对手在软件方案上缺乏对标CUDA的完整编程和工具链，而这些完整的工具链需要长时间积累，目前难以绕过CUDA去兼容英伟达的生态，所以CUDA成为了英伟达的坚实壁垒。
- 在PC时代，Wintel（Windows操作系统+英特尔芯片） 通过捆绑销售牢牢把握住对产业下游生产商的控制权，在消费端形成软硬结合的强大马太效应，微软和英特尔在市场获得巨大优势。在AI大模型时代，软硬件协同完善的生态是客户选择CUDA的原因，英伟达打造的GPU运算生态系统已拥有庞大的使用群体和客户粘性，英伟达算力王者的地位仍然稳健。

图：Wintel生态圈



图：CUDA生态系统

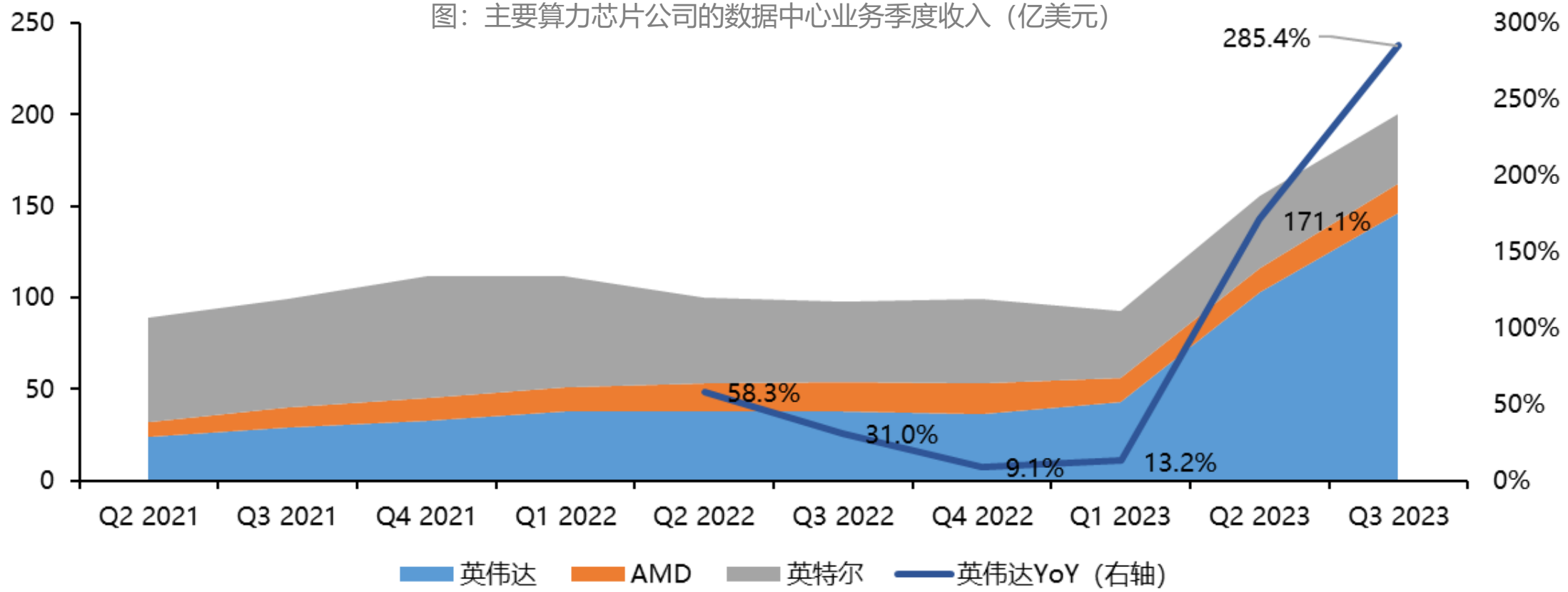


资料来源：英伟达，中航证券研究所

英伟达引领GPU计算潮流

- **英伟达“统治”AI算力芯片。**随着人工智能浪潮的兴起，给算力芯片行业带来了新的发展机会。英伟达的 GPU 在 AI 领域软硬件上都具有显著的优势和创新，可以为 AI 计算提供高性能、高并行度、高可扩展性和高能效等特点，满足大规模数据处理和复杂模型训练的需求。因此，英伟达的GPU主导了 AI 计算的市场。自2022年底，ChatGPT引爆算力需求以来，英伟达在数据中心的业务收入实现爆发式增长，2023Q2、Q3收入同比增速分别高达171%、285%。

图：主要算力芯片公司的数据中心业务季度收入（亿美元）



资料来源：ifind，中航证券研究所

英伟达芯片进口受限，国产算力需求迫切

- **围绕人工智能，美国已发起又一轮针对中国的制裁。**2022年9月，美国芯片巨头英伟达收到美国官方通知，若对中国和俄罗斯的客户出口两款高端GPU芯片——A100和H100，需要新的出口许可。2023年10月，美国的限制进一步增强，英伟达针对中国市场推出的“特供版” A800 和 H800 芯片遭到出口合规限制。美国商务部长雷蒙多表示，管制目的就是遏制中国获得先进芯片，从而阻碍中国「人工智能和复杂计算机领域的突破」。
- 国家层面，如果算力跟不上，则无法进行AI的大规模训练，将在人工智能竞争中落于下风。产业层面，算力的充足与否，直接决定了拥有AI技术和产品的公司，能否提供长期稳定的服务，从而在这场竞争中拥有先发权。随着一系列英伟达算力供给缩水动作，使得国内市场上已有的英伟达系列显卡进一步稀缺，算力一个月内涨价50%甚至100%已是常态。**因此，AI算力自主化对于国际科技竞争以及国内AI产业健康发展，都至关重要。**

图：美国商务部最新禁令

FOR IMMEDIATE RELEASE
October 17, 2023
https://bis.doc.gov

BUREAU OF INDUSTRY AND SECURITY
Office of Congressional and Public Affairs
Media Contact: OCPA@bis.doc.gov

Commerce Strengthens Restrictions on Advanced Computing Semiconductors, Semiconductor Manufacturing Equipment, and Supercomputing Items to Countries of Concern

Updates to Modify and Reinforce Restrictions Initially Released on October 7, 2022, to Address National Security Concerns Posed by PRC Military Modernization

Washington, D.C.--Today, the U.S. Department of Commerce's Bureau of Industry and Security (BIS) released a package of rules designed to update export controls on advanced computing semiconductors and semiconductor manufacturing equipment, as well as items that support supercomputing applications and end-uses, to arms embargoed countries, including the PRC, and to place additional related entities in the PRC on the Entity List.

资料来源：美国商务部，中航证券研究所

国产算力迎时代机遇

- 头部算力公司中，英伟达的强劲业绩增长反映了从通用到加速计算和生成式人工智能的广泛行业平台转型。NVIDIA GPU、CPU、网络、AI 代工服务和 NVIDIA AI Enterprise 软件都是全速增长的引擎。因此，英伟达被市场高度认可，2023年至今上涨230.8%，盈利和估值水平也显著优于竞争对手。
- 国产算力上市公司和海外公司技术和产品差距较大，并且在业绩端未能体现AI浪潮带来的增长，但由于人工智能的繁荣得到市场广泛押注，国内上市公司在估值端享受了本轮AI红利，股价强势增长。**我们认为，人工智能竞赛已经上升至大国竞争的高度，海外高端算力芯片被美国限制向中国出口，国内AI训练公司转向国产算力有望形成长期趋势，所以国产算力将迎发展机遇。**

图：算力公司市场表现情况（截至2023年12月15日）

	2023年度涨跌幅	市销率（TTM）	市盈率（TTM）	2023前三季度收入同比
英伟达	230.8%	26.62	63.23	85.5%
AMD	113.1%	10.08	1,071.82	-8.3%
英特尔	70.9%	3.60	-115.86	-20.8%
海光信息	84.0%	32.70	162.98	3.2%
寒武纪	165.4%	98.81	-53.88	-44.8%
龙芯中科	39.1%	73.40	-208.82	-18.5%

资料来源：ifind，中航证券研究所

国产算力蓬勃发展

- **中国 AI芯片产业逐渐呈现出蓬勃发展态势。**据IDC，2022年应用在数据中心的智能芯片数量超过百万个，其中本土品牌AI芯片数量已经接近15%的占比，涵盖品牌超过十余家。
- 国际科技网络巨头公司谷歌、脸书，亚马逊等等在AI芯片领域从云端训练到终端产品应用，在开源框架赋能产业行业上有一定的领先优势。国内企业也在打造从AI芯片注重云端训练+AI芯片终端响应+AI算法框架开源的生态体系。

表：英伟达国产替代产品

	产品型号	工艺制程	定位	理论对标英伟达产品
海思	异腾910	7nm	高端 台积电	英伟达A100/A800
	异腾310	12nm	中低端	
阿里	倚天710	5nm	高端	—
	含光800	12nm	中低端	英伟达P4
百度	昆仑芯2代	7nm	高端	—
腾讯	紫霄	12nm	中低端	英伟达A10
海光信息	深算二号	—	—	—
壁仞科技	BR100	7nm	中高端	
摩尔线程	MTT S3000	7nm		
燧原科技	云燧T20	12nm	中低端	
寒武纪	Gaudi	7nm		
沐曦	曦思N系列			
	曦云C系列			
景嘉微	JM9			

资料来源：IDC，中航证券研究所

目录

GPU行业需跨越CUDA横亘

- 英伟达CUDA已经实现了与算法工程师、芯片客户的强绑定，众多算法工程师已经习惯了一套工具库、一套编程语言，向外迁移存在不习惯等问题。所以很多算力芯片硬件厂商选择了兼容CUDA的路线——使硬件能够直接用CUDA调动起来，以降低用户的硬件迁移痛点。
- **兼容CUDA需要巨大时间和成本投入。**据集微网，兼容CUDA涉及50个驱动、50个编译器、50个数学库、300个应用层工程师，3-5年的时间。功能的验证，用户的培养需要额外3-5年，每年还要至少开支1000万-3000万元资助外部开发者。
- 从头部AI厂商布局来看，英伟达竞争者AMD选择兼容CUDA+自研原生“两条腿”并行，英伟达客户谷歌、Meta、亚马逊等均已推出自己的AI芯片。**我们认为，国产GPU在起步阶段兼容CUDA生态更容易发展，易于生存。在美国技术封锁的大背景之下，“AI信创”为国产算力芯片提供了市场窗口，远期国产GPU还是需要发展原生生态。**

图：头部厂商自研AI芯片

	公司	芯片名	发布时间	代数	制程（nm）
对标英伟达GPU	亚马逊	Trainium	2022	1	5
	亚马逊	Inferentia	2019	2	5
	Google	Tensor (TPU)	2015	4	7
	微软	Athena	2024	1	5
对标英特尔CPU	亚马逊	Graviton	2018	3	5
	Google	Maple	2025	1	5
	Google	Cypress	2025	1	5
	微软	Cascade	2024	1	5

资料来源：集微网、The information、中航证券研究所

国产算力有待提升

- 当下诸多国内本土芯片技术储备和生态能力仍围绕小模型时代的识别式人工智能展开，难以匹配大模型和生成式人工智能发展所需的软件生态、模型框架、性能需求，因此本土人工智能芯片仍需在发展、继承和竞争中成长。在中美脱钩的时代背景下，国产算力芯片正经历“可用”到“好用”的阶段，国产算力芯片整体实力有待提升。
- **从生态成熟度来看，我们认为，华为海思和海光信息有望率先替代英伟达算力芯片。**受益于华为ICT行业的领先地位，昇腾系列将获得从芯片设计、芯片制造、算力部署、应用生态的全方位支持，有更大概率成为主流国产算力芯片。

图：国内GPU与英伟达性能对比

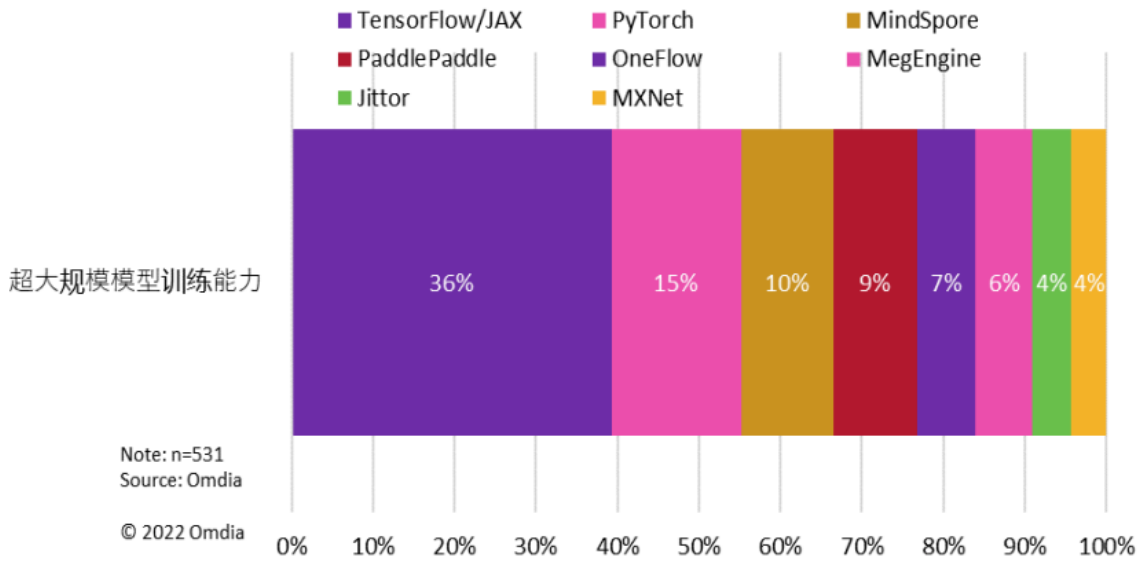
	英伟达	英伟达	华为海思	海光信息	寒武纪	摩尔线程
代表产品	A100	H100	昇腾910	深算二号	MLU290-M5	MTT S3000
FP32算力（TFLOPS）	19.5	60	-	-	-	15.2
FP16算力（TFLOPS）	624	2000	280	-	-	-
显存大小(GB)	80	80	32	-	32	32
显存带宽(TB/S)	2	3	-	-	1	0.4
互联带宽（GB/S）	600	900	56.5	-	600	-
功耗（W）	400	700	300	-	350	250
生态	CUDA	CUDA	昇思Mindspore	ROCm，兼容CUDA	-	-

资料来源：各公司官网，中航证券研究所

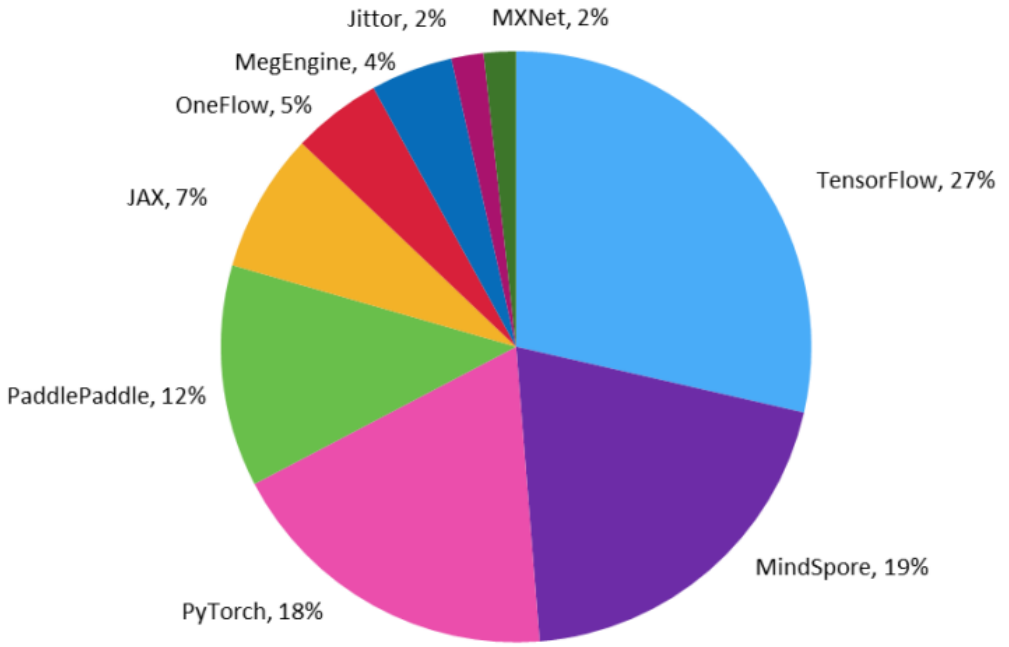
华为：Mindspore开发生态崛起

- 据Omdia，在支持超大规模模型训练开发方面,全球领先的人工智能框架TensorFlow和PyTorch仍然占据中国领导地位；**中国人工智能开发者认为,昇思MindSpore原生支持预训练大模型开发,已在中国市场上占据优势地位,并已经孵化出了一系列创新大模型。**
- 在以ChatGPT为代表的AIGC火爆的的背后,也出现了“造假”等AI伦理道德问题,人工智能开发者和机构越来越关注“负责任的人工智能”。Omdia在对人工智能开发者的调研中发现,在所有主流人工智能框架中,TensorFlow与MindSpore对“负责任的人工智能”提供的支持能力最好,分别位居第一与第二名。

图：TensorFlow, PyTorch, MindSpore和PaddlePaddle在支持超大规模模型训练方面处于领先



图：TensorFlow和MindSpore对“负责任的人工智能”提供的支持能力最好

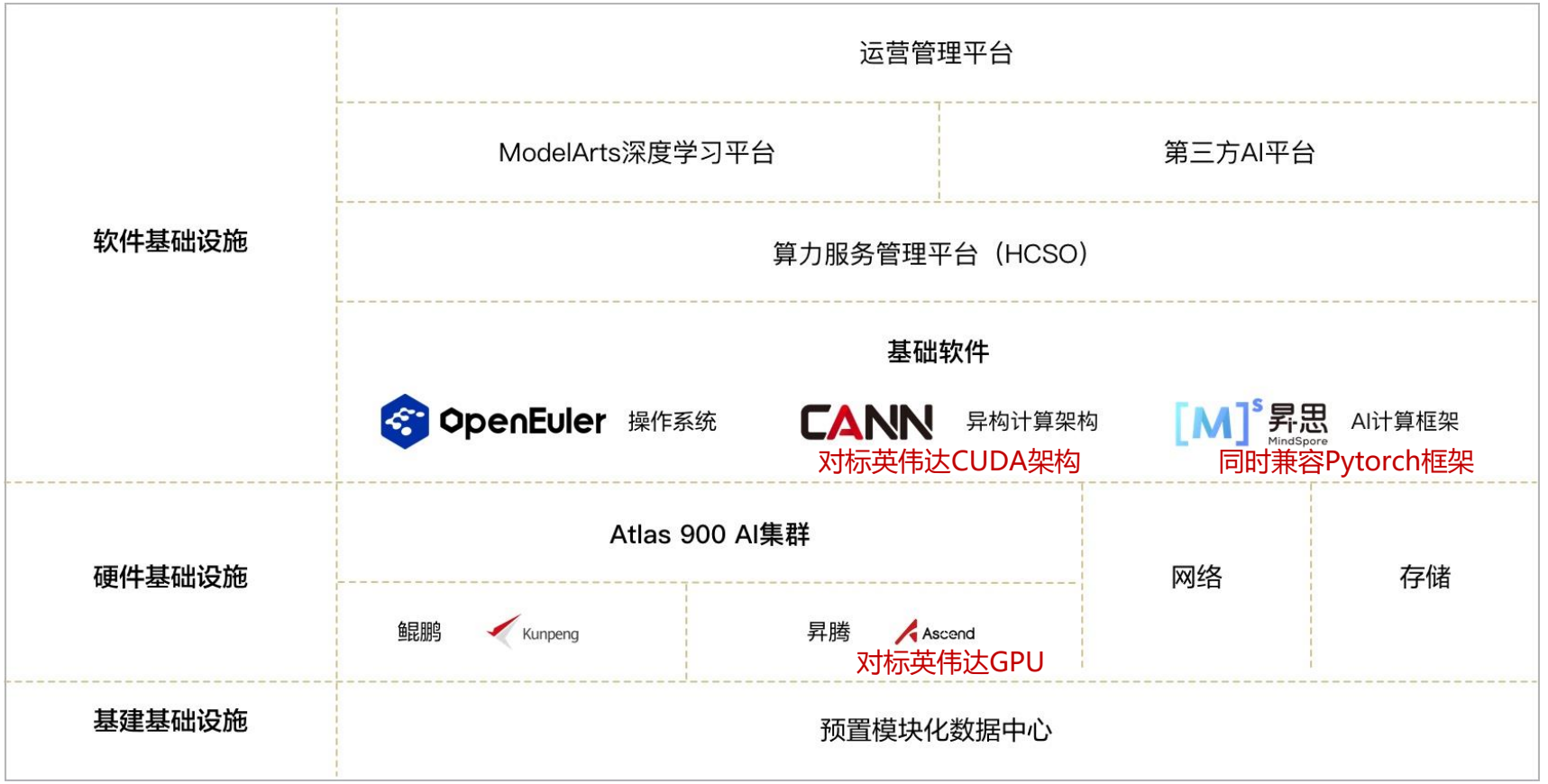


资料来源：Omdia《中国人工智能框架市场调研报告》、中航证券研究所

华为：打造全栈 AI 软硬件平台，构筑国产智能基石

- **华为通过芯片、异构计算架构、AI框架、AI开发平台等根技术的持续创新突破，打造自主的计算生态。**硬件方面，昇腾GPU对标英伟达GPU，软件方面；软件方面，CANN计算架构对标英伟达CUDA架构；昇思计算框架对标TensorFlow、Pytorch计算框架，并且PyTorch已同步支持昇腾NPU；应用方面，华为昇腾已经支持了业界50多个大模型，使能各行各业的智能化升级。

图：华为全栈 AI 软硬件平台



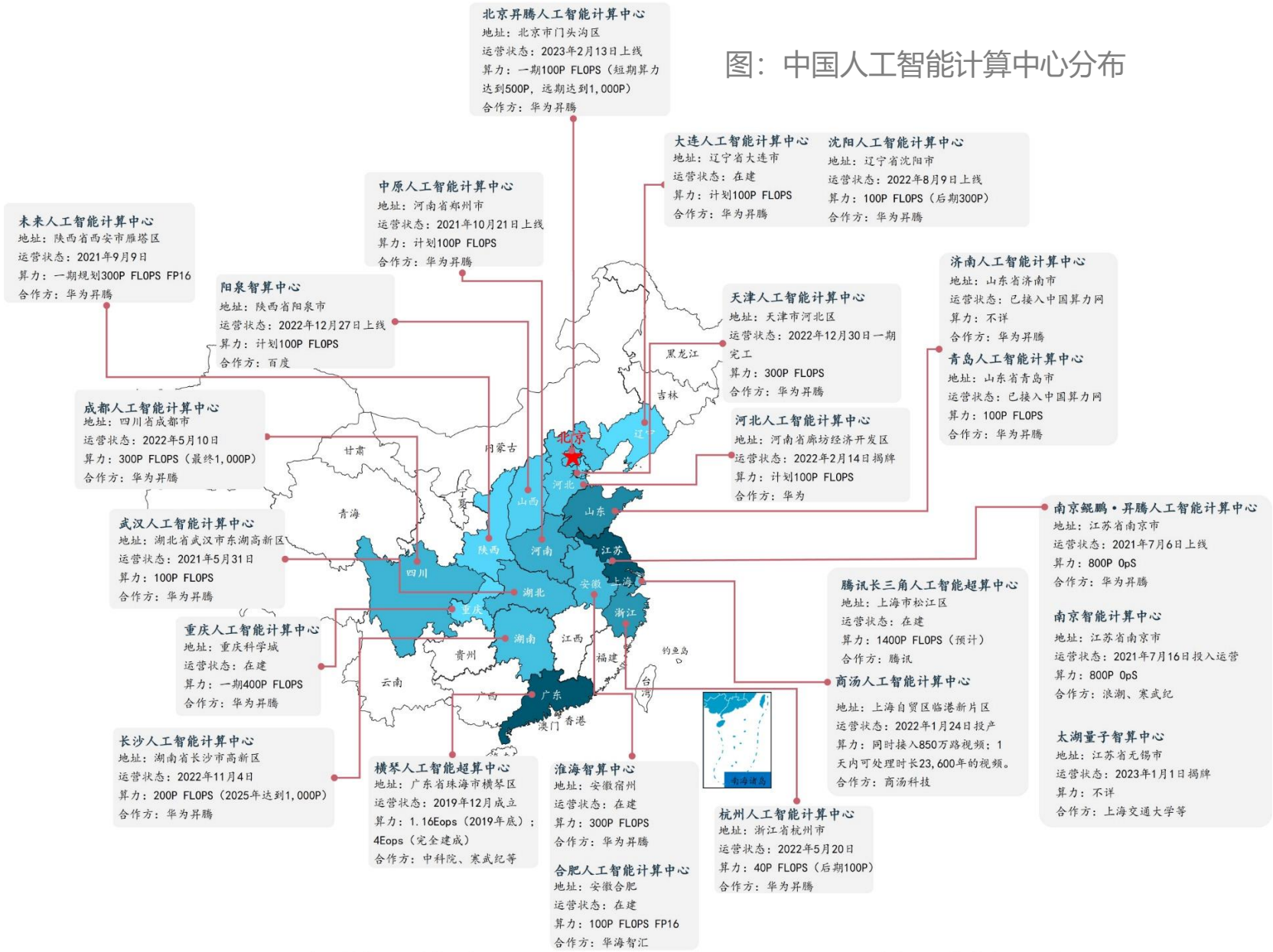
资料来源：华为、中航证券研究所

华为：引领国内AI芯片发展

- 2023年1月，国家工业信息安全发展研究中心提出，我国智能算力中心建设已进入2.0阶段，在智能计算中心的建设过程中，投-建-运往往由不同的单位负责，导致资源浪费，体验不佳。**所以2.0阶段，算力中心的建设、运营、应用应该联动，满足一体化服务需求。**

- **华为AI芯片在国内优势显著，因为华为在ICT领域具备从建设到运营的全环节壁垒。**至顶智库统计，截至2023年2月，我国目前已投入运营和在建的人工智能计算中心达23个，有16家人工智能计算中心由华为参与建设，使用华为昇腾人工智能计算系统作为底层技术支持。国内约70%算力中心由华为参与，叠加国产AI芯片自主化的需求，所以华为AI芯片将长期引领国内算力芯片发展。

图：中国人工智能计算中心分布



资料来源：至顶网、中航证券研究所

海光信息：CPU+GPU布局完整，产品迭代顺利

- 海光信息的产品包括海光通用处理器（CPU）和海光协处理器（DCU）。公司是少数几家同时具备高端通用处理器和协处理器研发能力的集成电路设计企业。基于x86指令框架、“类CUDA”计算环境和国际先进处理器设计技术，公司大力发展满足中国信息化发展需要的高端处理器产品。
- **强大研发实力推动公司产品持续迭代。**公司骨干研发人员多拥有国内外知名芯片公司的就职背景，拥有成功研发x86处理器或ARM处理器的经验。公司2021全年、2022全年、2023前三季度的研发投入分别为7.45亿元、14.14亿元、12.79亿元，迅速增长的研发投入，支撑公司产品不断迭代。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，深算二号于2023Q3发布，深算三号研发进展顺利。

图：海光信息主要产品

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

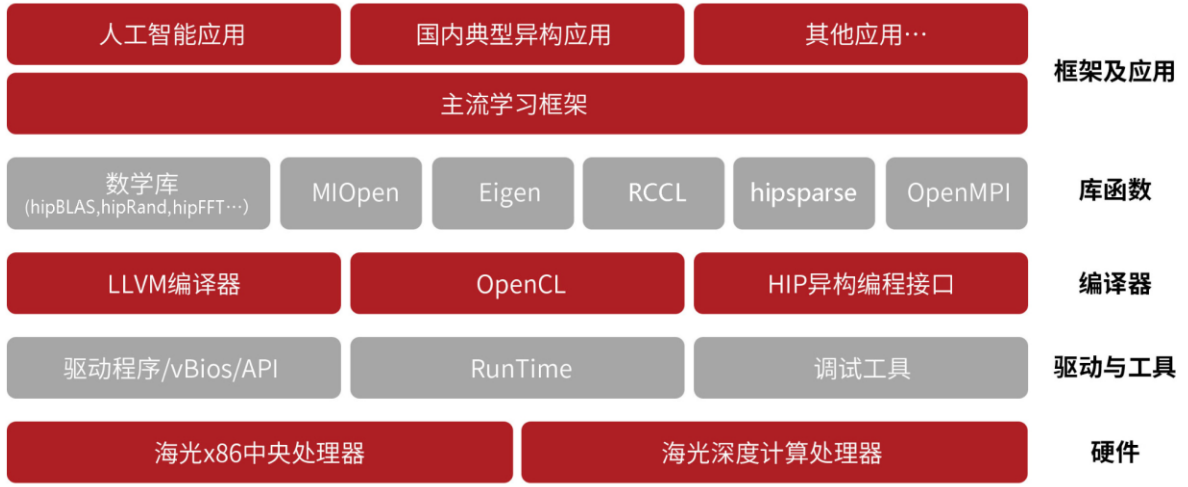
资料来源：海光信息招股书、中航证券研究所

海光信息：兼容“类CUDA“生态，打造全栈AI基础设施

- 海光DCU属于GPGPU的一种，全面兼容ROCm GPU计算生态。ROCm是AMD的软件平台，用来加速 GPU 计算，对标英伟达的CUDA平台。AMD的GPU上编程模型使用的是HIP或者OpenCL，运行环境是ROCm。英伟达显卡上，编程模型是CUDA，运行环境也是CUDA。
- 由于ROCm和CUDA在生态、编程环境等方面具有高度的相似性，CUDA用户可以以较低代价快速迁移至ROCm平台，ROCm也被称为“类CUDA”。因此，海光DCU协处理器能够较好地适配、适应国际主流商业计算软件和人工智能软件。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。

图：公司OEM客户

图：海光提供完善的AI软件栈支持



资料来源：海光信息、中航证券研究所

寒武纪：云边端AI硬件协同发展

- 寒武纪提供云边端一体、软硬件协同、训练推理融合、具备统一生态的系列化智能芯片产品和平台化基础系统软件。
- 公司云端产品线前五大客户，包括中科可控、浪潮信息等服务器厂商及阿里、百度等互联网公司，目前公司已完成了服务器厂商的产品适配并与主要服务器厂商建立了长期深入的合作关系。

图：寒武纪芯片产品面向云、边、端三大场景

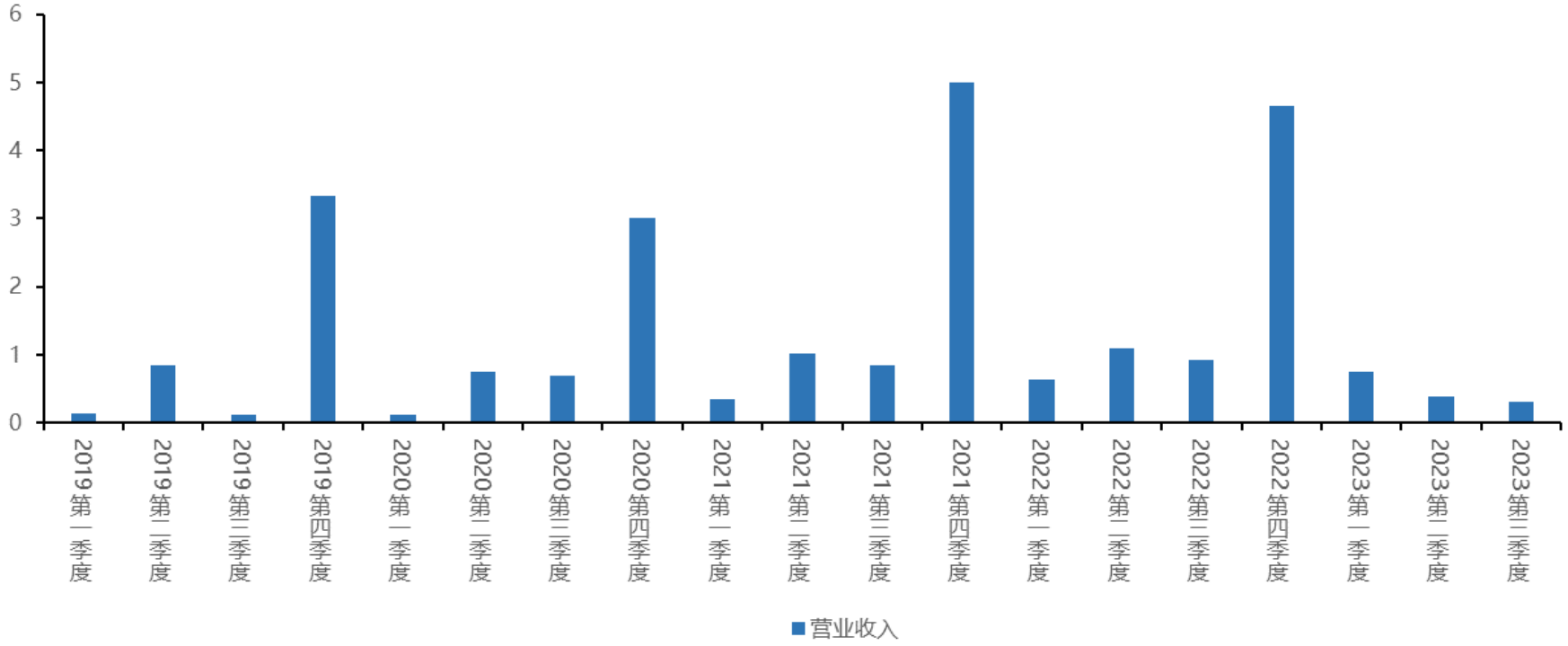
产品线	产品类型	寒武纪主要产品	推出时间
云端产品线	云端智能芯片及加速卡	思元100（MLU100）芯片及云端智能加速卡	2018年
		思元270（MLU270）芯片及云端智能加速卡	2019年
		思元290（MLU290）芯片及云端智能加速卡	2020年
		思元370（MLU370）芯片及云端智能加速卡	2021年、2022年
	训练整机	玄思1000智能加速器	2020年
		玄思1001智能加速器	2022年
边缘产品线	边缘智能芯片及加速卡	思元220（MLU220）芯片及边缘智能加速卡	2019年
IP授权及软件	终端智能处理器IP	寒武纪1A处理器	2016年
		寒武纪1H处理器	2017年
		寒武纪1M处理器	2018年
	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、中航证券研究所

寒武纪：互联网客户存较大增长空间

- 公司四季度收入较前三季度大幅增长，是由于智能计算集群系统业务主要服务于城市智能计算中心客户，该类客户第四季度组织项目进度评审及项目验收工作。从公司收入季节性明显的特征分析，公司目前客户可能主要为回款较慢的政府、学校、金融等G端客户，公司在互联网端存较大增长空间。

图：公司单季度收入（亿元）



资料来源：ifind、中航证券研究所

目录

AI算力刺激更多存储器用量

- **AI服务器存储容量倍增，带动存储器需求成长。**据TrendForce，AI服务器需要配置更多DRAM、SSD和HBM等大容量存储以应对日益复杂的大模型所带来的海量数据。当前普通服务器DRAM普遍配置约为500至600GB，而AI服务器DRAM 配置可达1.2至1.7TB，是普通服务器的二到三倍。此外，相较于一般服务器而言，AI服务器多增加GPGPU的使用，因此以NVIDIA A100 80GB配置4或8张计算，HBM用量约为320~640GB。未来在AI模型逐渐复杂化的趋势下，将刺激更多的存储器用量，并同步带动 Server DRAM、SSD 以及 HBM 的需求成长。

图：AI服务器 vs 普通服务器存储用量

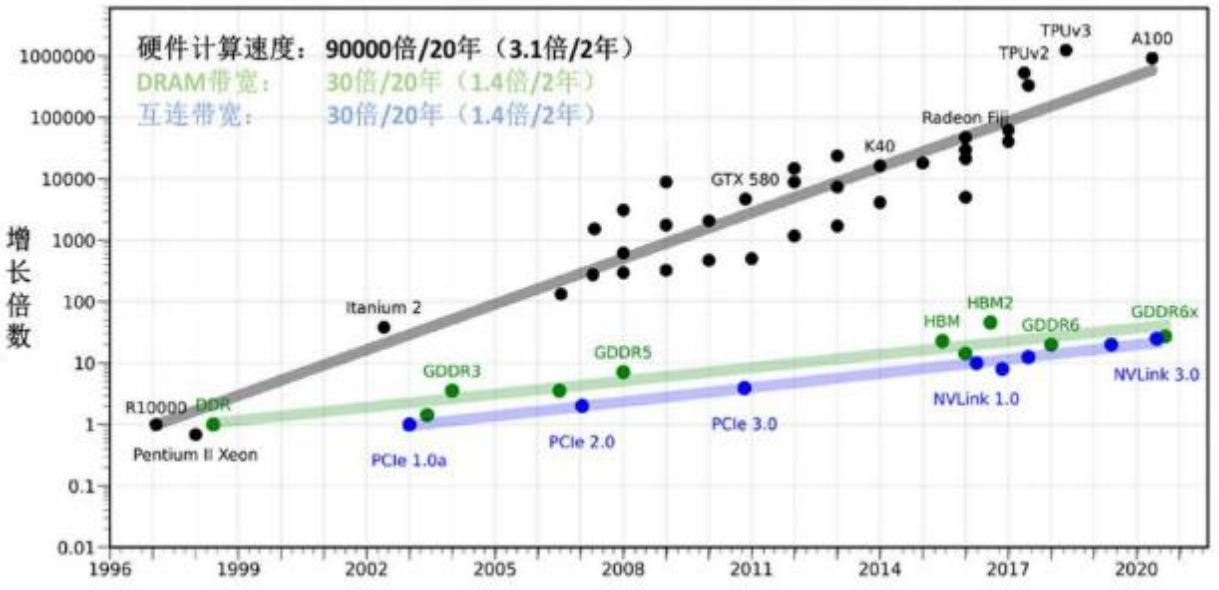
	普通服务器	AI服务器	未来AI服务器
DRAM容量	500-600GB	1.2-1.7TB	2.2-2.7TB
SSD容量	4.1TB	4.1TB	8TB
HBM容量	——	320-640GB	512-1024GB

资料来源： TrendForce、中航证券研究所

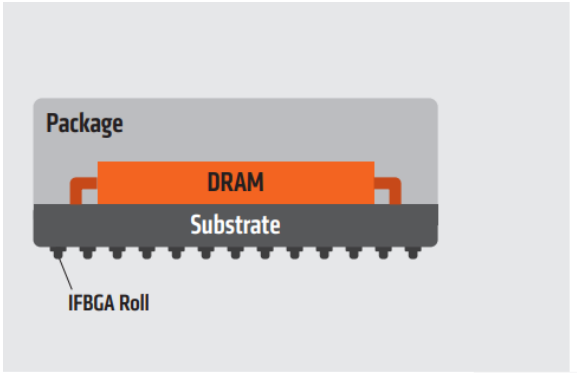
HBM解决GPU内存危机

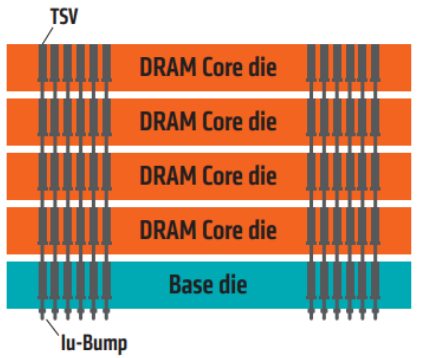
- 处理器的性能按照摩尔定律规划的路线不断飙升，内存所使用的DRAM却从工艺演进中获益很少，性能提升速度远慢于处理器速度，造成了DRAM的性能成为制约计算机性能的一个重要瓶颈，即所谓的“内存墙”。**HBM成为增加存储器带宽的路径之一，以解决大数据时代下的“内存墙”问题。**
- HBM（High Bandwidth Memory）即高带宽存储器，按照JEDEC的分类，HBM属于GDDR内存的一种，其通过使用先进的封装方法（如TSV硅通孔技术）垂直堆叠多个DRAM，并与GPU封装在一起。HBM主要优势是在高带宽和低功耗领域，应用场景以配合并行计算的GPU和ASIC芯片为主。

图：存储计算性能存在“剪刀差”



图：HBM性能优势明显





GDDR5	Per Package	HBM
32-bit	Bus Width	1024-bit
Up to 1750MHz (7GBps)	Clock Speed	Up to 500MHz (1GBps)
Up to 28GB/s per chip	Bandwidth	>100GB/s per stack
1.5V	Voltage	1.3V

资料来源：中国移动研究院、AMD、中航证券研究所

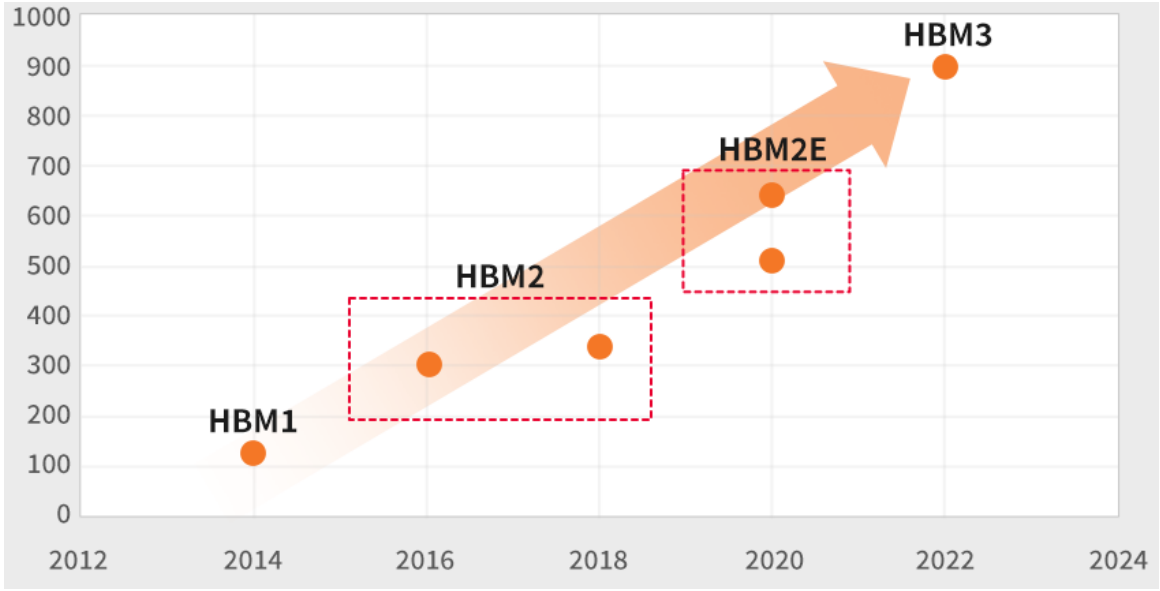
HBM是GPU主流内存方案

- **主流用于大模型训练的AI芯片，其显存方案跟随HBM技术的提升而演进，向更高带宽的方向发展。**英伟达的A100和AMD MI200均采用HBM2e方案，而英伟达最新一代的H100芯片则采用SK海力士的HBM3方案，AMD发布的MI300X采用192GB的HBM3内存方案，带宽最高可达H100的1.6倍。

图：部分AI训练芯片内存方案

GPU型号	推出时间	存储技术	GPU 显存容量	显存带宽
NVIDIA V100 SXM2	2017年	HBM2	16GB/32GB	900 GB/s
NVIDIA A100 80GB SXM	2020年	HBM2e	80GB	2039GB/s
NVIDIA H100 SXM	2022年	HBM3	80GB	3.35TB/s
AMD Instinct MI100	2020年	HBM2	32GB	1.2 TB/s
AMD Instinct MI200	2021年	HBM2e	128GB	3.2TB/s
AMD Instinct MI300X	2023年	HBM3	192GB	5.2TB/s

图：HBM约两年迭代一次带宽（GB/s）

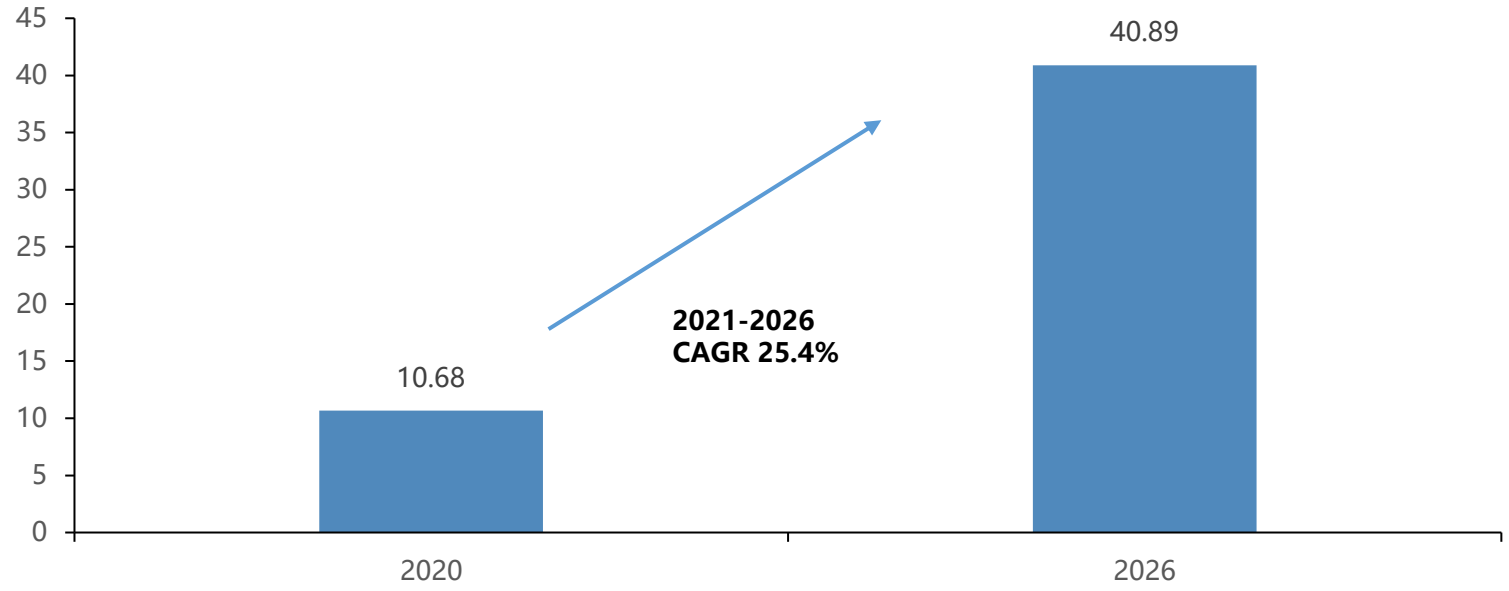


资料来源： NVIDIA官网、AMD官网、奎芯科技、海力士、中航证券研究所

HBM供不应求

- **AI创造纯增量市场，HBM需求量年增近六成。**目前高端AI服务器GPU搭载HBM已成主流，TrendForce预估2023年全球HBM需求量将年增近六成，来到2.9亿GB，2024年将再成长三成。根据Mordor Intelligence，2020年HBM市场价值为 10.68 亿美元，预计到2026年将达到40.89亿美元，在2021-2026年预测期间的复合年增长率为25.4%。HBM头部企业SK海力士在2023年7月表示，目前其HBM的销量占比还不足营收1%，但2023年销售额占比有望成长到10%，同时预计在2024年应用于AI服务器的HBM和DDR5的销量将翻一番。2023年10月，SK海力士表示，已经在2023年出售了明年HBM3和HBM3E的所有产量。

图：HBM市场规模预测（亿美元）

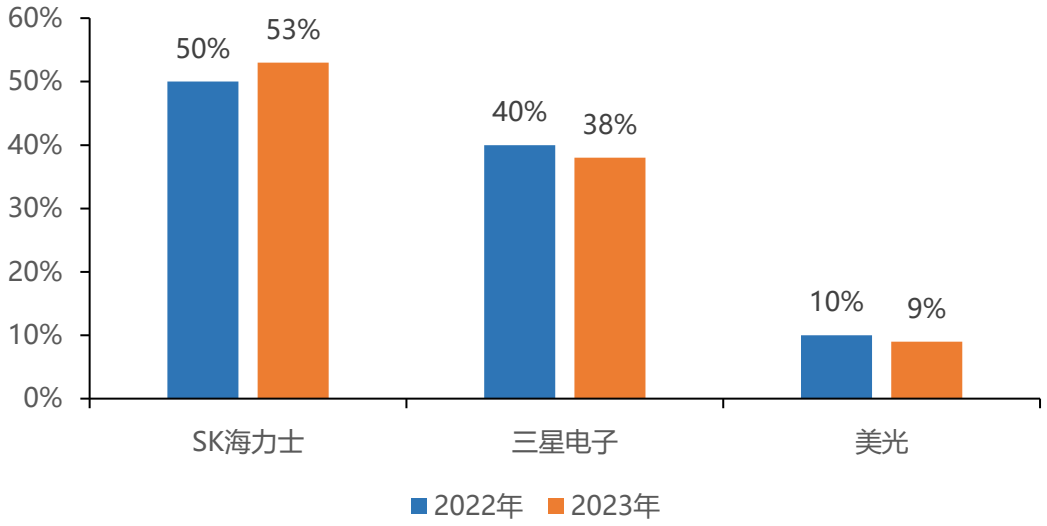


资料来源： Mordor Intelligence、中航证券研究所

HBM成为存储下一主战场，头部厂商相继扩产

- **HBM价值量显著敢于标准DRAM，成为新利润增长点。**芯片咨询公司 SemiAnalysis 表示，HBM的价格大约是标准DRAM芯片的五倍，为制造商带来了更大的总利润。目前，HBM占全球内存收入的比例不到5%，但SemiAnalysis预计到2026年将占到总收入的20%以上。SK海力士首席财务官Kim Woo-hyun 在4月份的财报电话会议上表示预计2023年HBM收入将同比增长50%以上。
- **韩系存储供应商积极扩产，主导HBM市场。**根据TrendForce，2022年三大原厂HBM市占率分别为SK海力士约50%、三星约40%、美光约10%。高阶深度学习AI GPU的规格也在刺激HBM产品更迭，2023下半年伴随NVIDIA H100与AMD MI300的搭载，三大原厂也已规划相对应规格HBM3的量产。SK海力士作为目前唯一量产新世代HBM3产品的供应商，其整体市占率有望提升至53%，而三星、美光则预计陆续在2023年底至2024年初量产，市占率分别为38%及9%。**基于各原厂积极扩产的策略，HBM供需比有望获改善，预估将从2023年的-2.4%，转为0.6%。**

图：2022年及2023年HBM市场份额预测



图：三大存储厂的HBM开发进度

	Brand	Speed (Gbps)	Tech Nodes	2022				2023				2024				2025				2026			
				1Q22	2Q22	3Q22	4Q22	1Q23	2Q23	3Q23	4Q23	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26	2Q26	3Q26	4Q26
HBM2e	Samsung	3.2-3.6	1Y 16Gb	8/16GB														EOL Time : Not Fixed					
	SK hynix	3.6	1Y 16Gb	8/16GB																			
	Micron	3.2-3.6	1Z 16Gb	16GB							EOL												
HBM3	Samsung	6.4	1Z 16Gb					16GB		24GB													
	SK hynix	5.6-6	1Z 16Gb					16GB		24GB													
HBM3e	Samsung	8	1alpha 24Gb									24GB		36GB									
	SK hynix	8	1beta 24Gb									24GB			36GB								
	Micron	8	1beta 24Gb									24GB			36GB								
HBM4		TBD	TBD	Full spec may be released in 2H24-2025; C/S in 2026																			

Source: TrendForce, Aug., 2023

资料来源： TrendForce、中航证券研究所

建议关注：香农芯创——深度布局存储产业

- **拥抱头部原厂资源，投资半导体产业链协同赋能。**香农芯创目前拥有两大业务板块：电子元器件分销平台，和半导体产业链协同赋能。公司第一大供应商为SK海力士，并获得了SK海力士、MTK等原厂的授权代理权，形成了代理原厂线优势。公司投资半导体设计、封测、设备、应用等各个环节领军企业，推动半导体产业链生态发展和升级。
- **与海力士合作拓展SSD业务。** 5月26日，公司公告与深圳大普微电子、江苏聿泉君海荣芯投资等合作方共同出资设立深圳市海普存储科技有限公司。新公司拟开展SSD存储产品的设计、生产和销售业务，当前我国企业级SSD市场国产化率较低，而2022-2026年我国企业级SSD市场规模将以23.7%的年均复合增速成长，市场空间广阔。新公司的成立，助力打破技术垄断，推进企业级SSD国产替代进程。此外，投资方江苏聿泉君海荣芯投资的第一大股东为SK海力士（无锡）投资公司，利益共享或有技术赋能。

锁定一线品牌代理资格

SK hynix　MEDIATEK　Cambricon 寒武纪　GigaDevice 兆易创新

投资赋能，与半导体产业链协同发展

FHEC 甬矽电子 FOREHOPE ELECTRONIC　壁仞科技 BIREN TECHNOLOGY　HD 好达　Leadmicro 微导

资料来源： 香农芯创、中航证券研究所

建议关注：雅克科技——海力士重要供应商

- 雅克科技是平台型先进材料公司，形成了以电子材料为核心，LNG保温板材为补充的战略模式。公司收购海外优质资产弯道超车，并打入海外核心客户供应链，技术实力国内领先。公司半导体材料包括前驱体、光刻胶及辅助化学品、电子特气、硅微粉等，客户包括台积电、三星电子、Intel、中芯国际、海力士、京东方等业国际头部企业。
- 2023年9月，公司发布对外投资公告，子公司江苏雅克半导体以约2.7亿元的价格收购SKenpulse公司持有的SKC-ENF75.1%的股权，SKC-ENF持有爱思易（江苏）公司及爱思开希（南通）公司100%的股权，主要产品为半导体光刻胶辅助化学品，包括显影液、稀释剂、蚀刻液等，目前是国内唯一拥有相关技术的本土供应商，产品可以满足对应全品类光刻胶匹配使用。公司横向布局湿电子化学品，拓展业务疆界，并增强主业的规模效应。

前驱体

推出新High-K材料及超高/低温硅类产品，是海力士主要供应商

江苏先科工厂量产在即

硅微粉

球形氧化铝等产品已开始向客户稳定供货

亚微米球形二氧化硅完成研发，供货给住友电木、日立化成等知名环氧塑封料的厂商

特气

工业用六氟化硫受益于国内特高压输变电需求

子公司科美特供应台积电、三星电子、中芯国际、海力士等头部客户

光刻胶

OLED用低温RGB光刻胶、CMOS传感器用RGB光刻胶、先进封装RDL层用i-Line光刻胶等导入客户测试

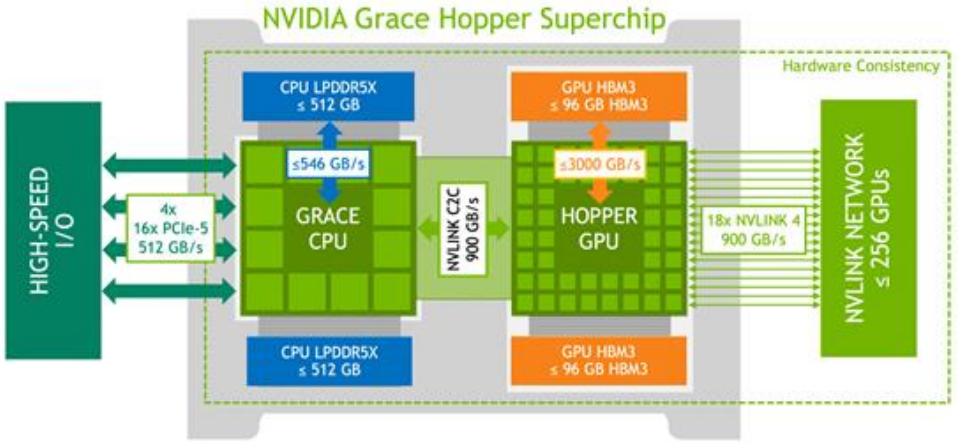
资料来源： 雅克科技、中航证券研究所

目录

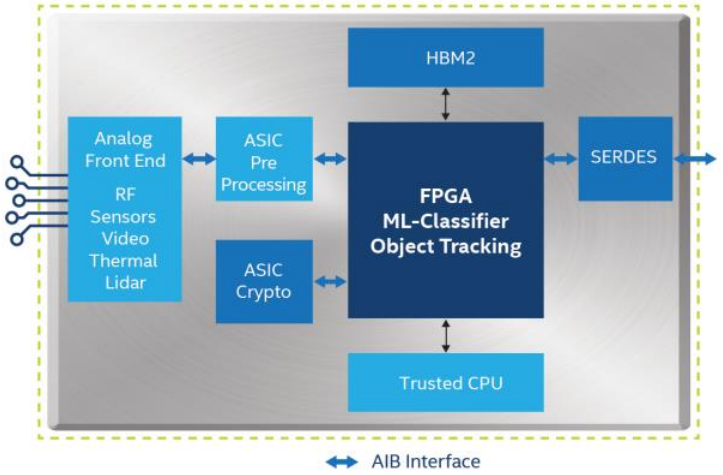
异构计算时代，先进封装战略地位凸显

- 异构集成（Heterogeneous Integration），即横向和纵向连接多个半导体，可将更多的晶体管装在一个更小的半导体上，准确地说是在更小的半导体封装内，从而提供比其各部分之和更大的功用。CPU+GPU是人工智能异构计算的主要组合形式，英伟达的Grace Hopper超级芯片通过异构集成CPU、GPU以及存储器，实现芯片更高带宽的互连，能够承担更大的数据集、更复杂的模型和新的工作负载。
- **先进封装成为突破“摩尔定律”局限的技术。**先进封装技术充当着半导体器件与系统之间的桥梁，是实现异构集成的关键技术，因此，这种连接方法变得越来越重要。先进封装技术本身已成为一种系统解决方案，半导体头部设计、制造商均通过此方法，在摩尔定律放缓的时代，从系统层面继续提升芯片性能。

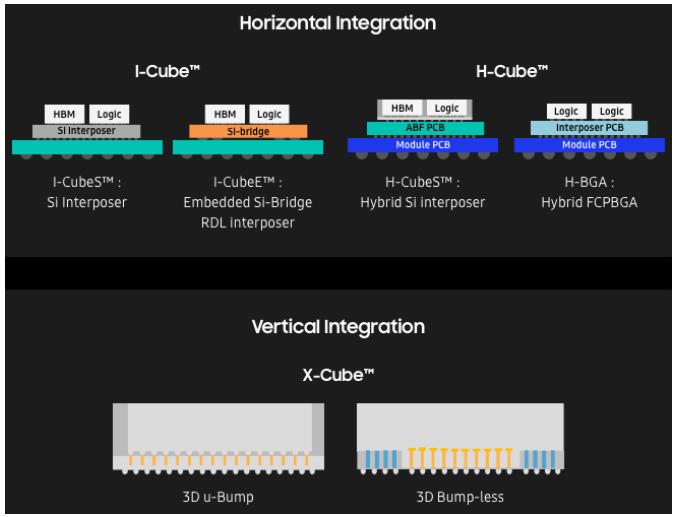
图：英伟达Grace Hopper 超级芯片集成GPU\CPU\存储



图：英特尔异构 3D 系统级封装集成



图：三星先进异构集成



资料来源： 英特尔、英伟达、三星、中航证券研究所

顺应AI算力芯片发展，封装技术持续创新

- **自2020年开始，先进封装进入技术融合期，成为关键的系统级解决方案。**不仅需要借助多项技术将各类芯片集成到同一封装内，还需要在整合系统时将多个部分连接至同一模块。AI计算芯片融合了多项先进封装技术，HBM应用TSV堆叠技术获得超高带宽，而为了将HBM和GPU集成，CoWoS封装技术被深度开发。因此，封装技术将成为提供整体系统解决方案的重要手段 。海力士判断，未来各公司将依赖封装技术助力其成为半导体行业的领军者。

图：封装行业发展的三个阶段

资料来源： 海力士、中航证券研究所

顺应AI算力芯片发展，封装技术持续创新

- 随着封装技术的发展，近十年中，重新分配层（RDL）、倒片封装（Flip Chip）和硅穿孔（TSV）等封装技术得到了积极广泛的应用，在硅晶圆或芯片堆叠结构晶圆中进行工艺处理，大幅提高了产品的性能和容量。SK海力士凭借业界领先的TSV堆叠技术引领了市场发展，这其中包括HBM封装存储器解决方案，以及用于服务器的高密度存储器（HDM）三维堆叠技术。同时，海力士持续迭代封装技术，研发了批量回流模制底部填充、混合键合、扇出型晶圆级封装等技术，以进一步提升了HBM的堆叠层数。

图：堆叠封装层数增长，提高HBM带宽

四代HBM规格比较（以SK海力士产品为例）				
类别	HBM1	HBM2	HBM2E	HBM3
带宽	128GB/s	307GB/s	460GB/s	819GB/s
堆叠高度	4层	4层/8层	4层/8层	4层/12层
容量	1GB	4GB/8GB	8GB/16GB	16GB/24GB
I/O速率	1Gbps	2.4Gbps	3.6Gbps	6.4Gbps

资料来源： 海力士、中航证券研究所

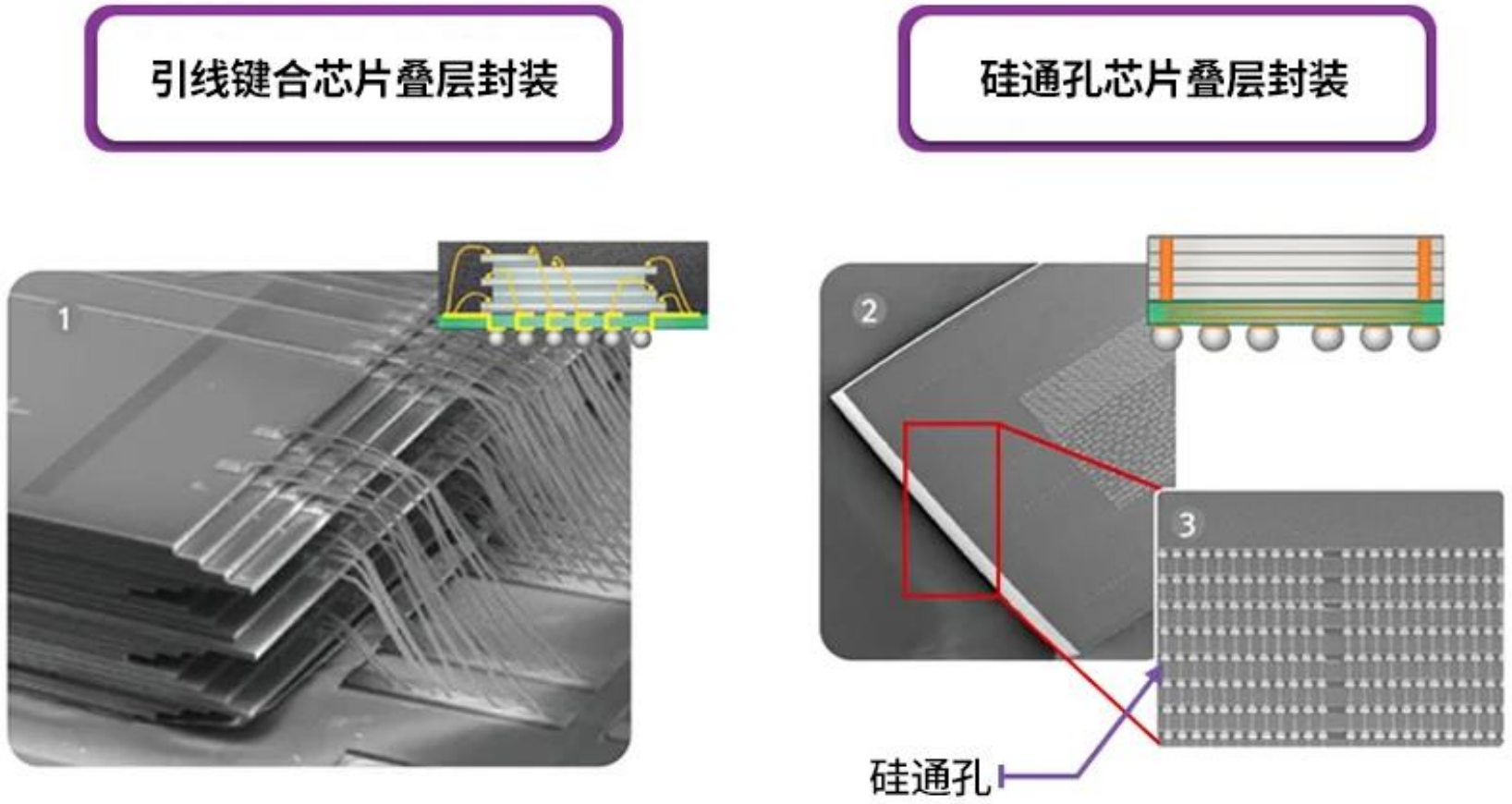
图：海力士持续迭代封装技术

硅通孔封装技术大幅提升HBM带宽

- HBM采用一种全新的DRAM架构，这种架构借助硅通孔技术（TSV）来增加引脚数量。通常，在DRAM规范中，“X4”表示有四个引脚用于发送信息，或可以同时从DRAM发送4位（bit）信息。相应地，X8表示8位，X16表示16位，以此类推。增加引脚数量有利于同时发送更多信息。

- 由于堆叠芯片以及连接引脚（Pin）的数量增加，引线变得更加复杂，而且也需要更多空间来容纳这些引线。相比之下，硅通孔芯片堆叠则不需要复杂的布线，因而封装尺寸更小。因此，由于自身局限性，引线芯片堆叠最多只能达到X32，而硅通孔堆叠则没有这方面的局限性，使HBM可达到X1024。

图：封装剖面图

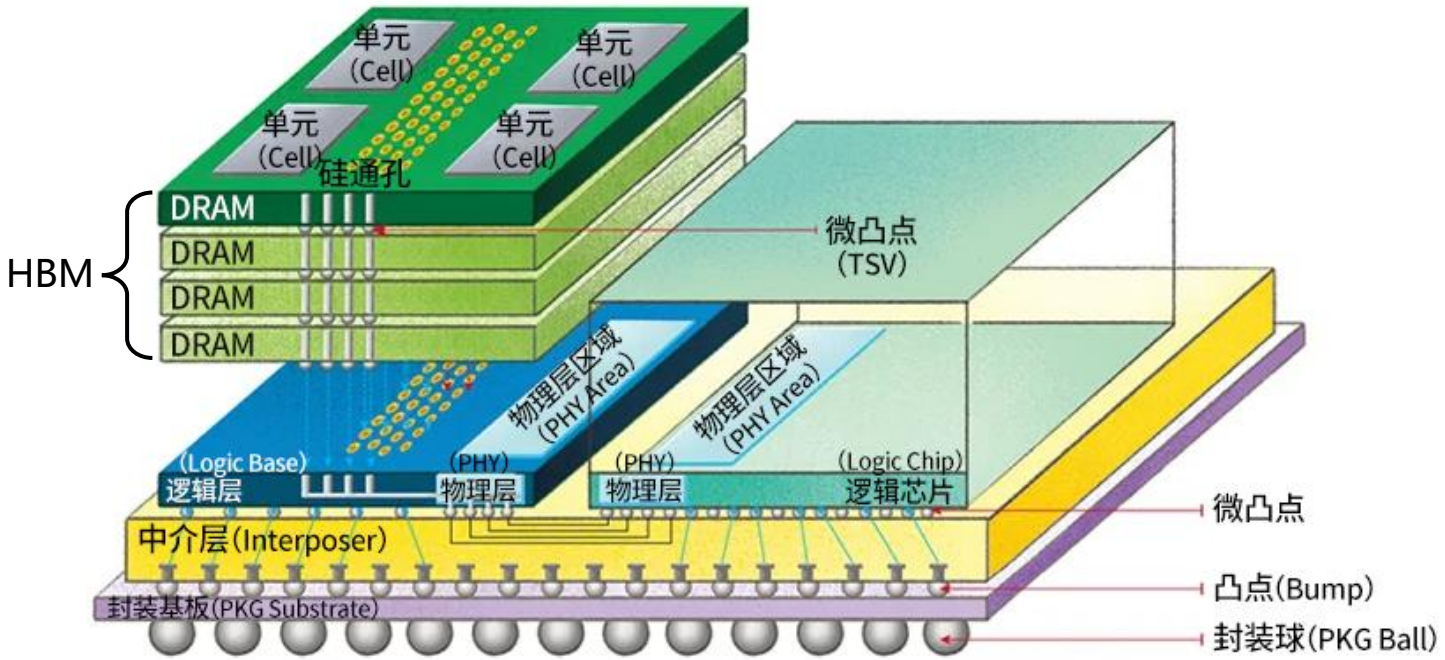


资料来源：海力士、中航证券研究所

台积电CoWoS封装集成HBM与GPU

- HBM并非一种全封装产品，而是一种半封装产品。当HBM产品被送到系统半导体制造商那里时，系统半导体制造商会使用中介层构建一个2.5D封装，将HBM与逻辑芯片并排排列。由于2.5D封装中的基板无法提供用于支持HBM和逻辑芯片的所有输入/输出引脚的焊盘（Pads），因此需要使用中介层来形成焊盘和金属布线，从而容纳HBM和逻辑芯片。然后，再将这些中介层与基板连接。
- **台积电CoWoS技术主导2.5D封装。**CoWoS是英伟达选择的主流封装技术，能够以合理的成本提供最高的互连密度和最大的封装尺寸。由于目前大部分HBM系统都封装在CoWoS上，因此，大部分先进的数据中心GPU由台积电在CoWoS上封装。

图：使用HBM的2.5D封装

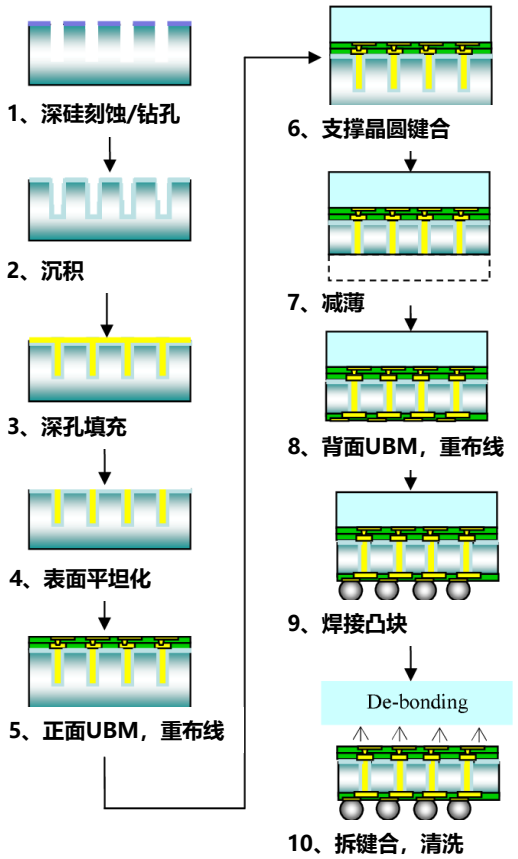


资料来源： 海力士、中航证券研究所

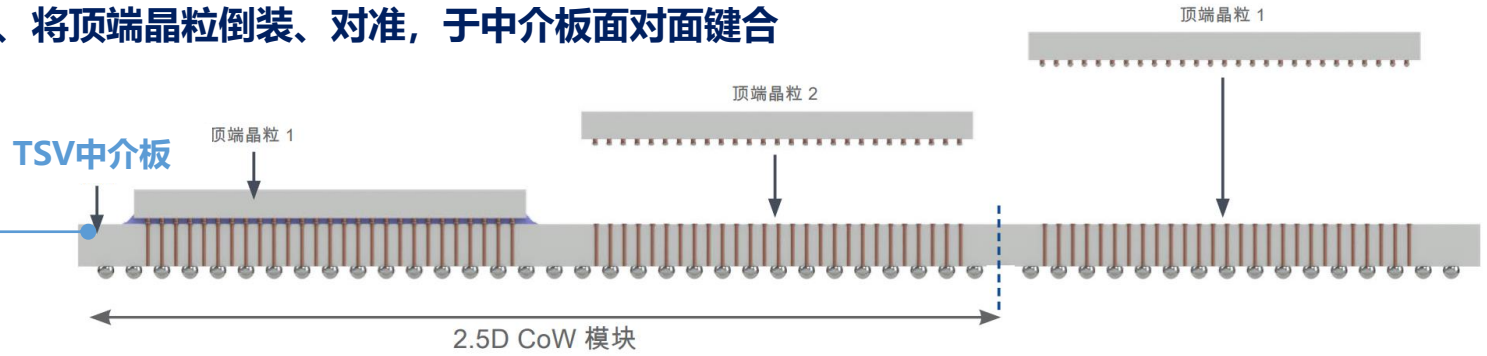
CoWoS封装流程

■ CoWoS由 CoW 和 oS 组合而来。 先将芯片通过 Chip on Wafer（CoW）的封装连接至中介板，再把 CoW 模块与基板（Substrate）连接。

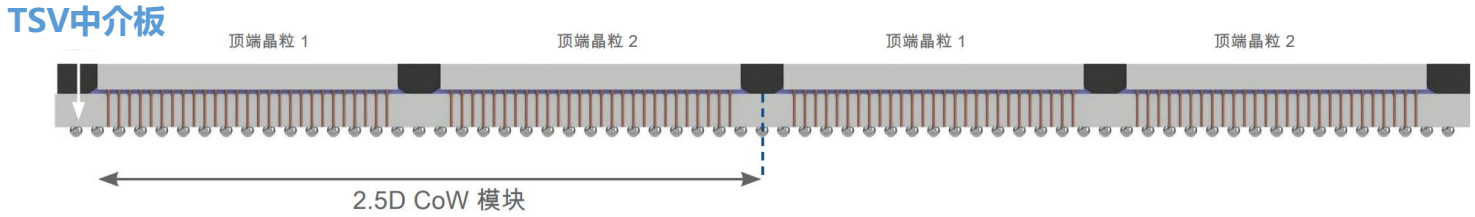
TSV中介板制作流程



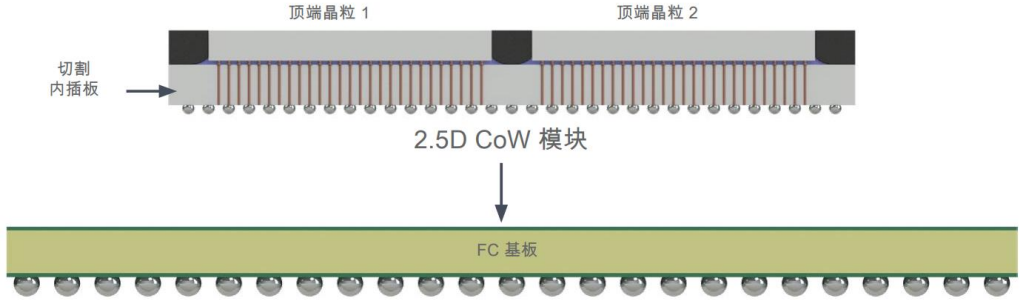
1、将顶端晶粒倒装、对准，于中介板面对面键合



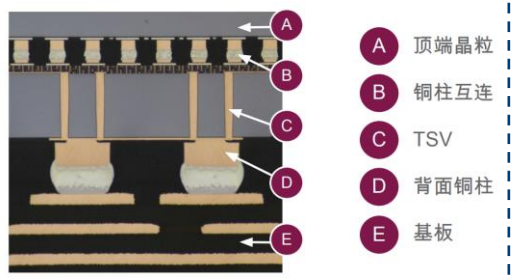
2、晶圆级包覆成型、塑封晶圆研磨和模块减薄



3、切割CoW模块，通过凸块与封装基板相连（CoW+oS）



完工的封装晶圆

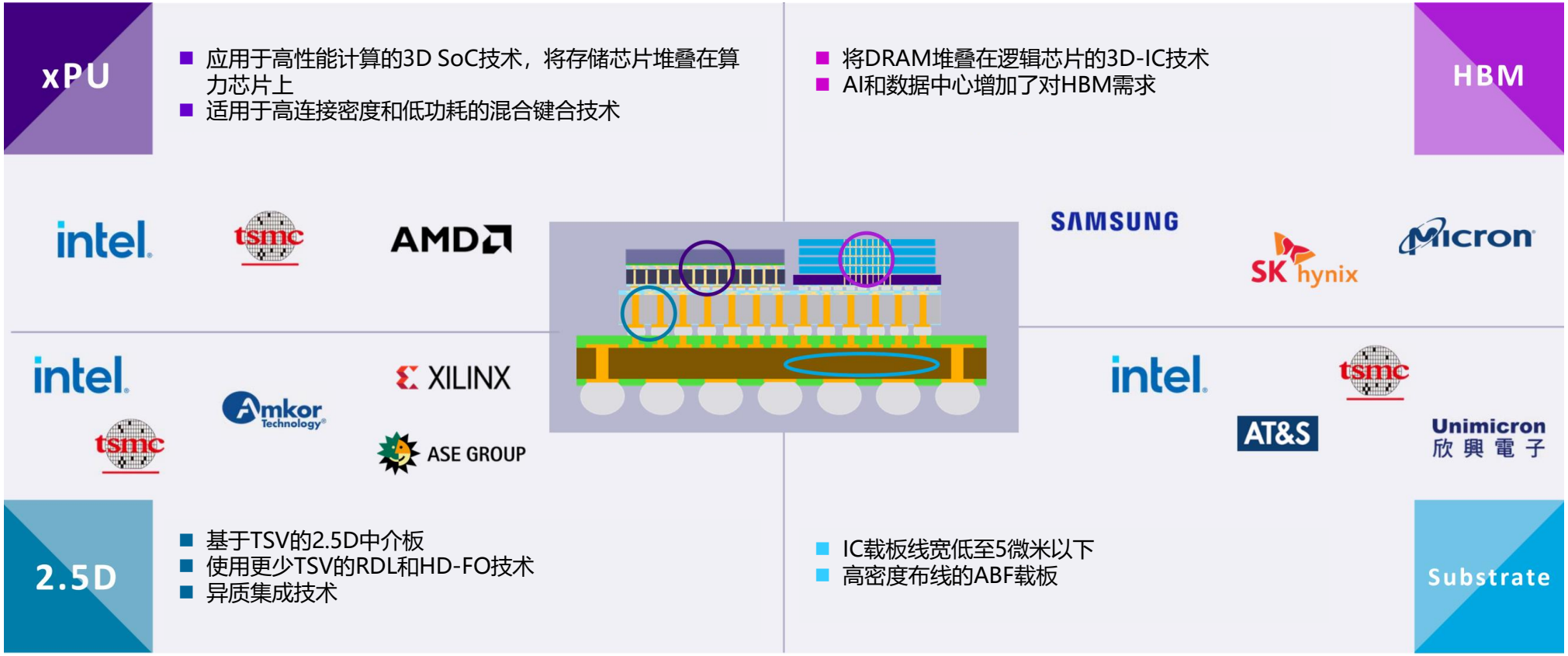


资料来源： Amkor、《TSV interposer fabrication for 3D IC packaging》,中航证券研究所

先进封装生态已形成，国产替代空间广阔

- 先进封装生态涵盖从芯片设计、制造、材料的供应商。包括高性能算力芯片巨头英特尔、英伟达、AMD；存储芯片供应商三星、海力士、镁光；先进封装工艺服务商台积电、英特尔、日月光；IC载板供应商欣兴电子、英特尔、AT&S等。
- 我国本土供应商在先进封装产业链的参与度较低，在逆全球化的背景下，除了实现高阶芯片制程的自主可控，先进封装的国产化也同样迫在眉睫。

图：先进封装产业生态

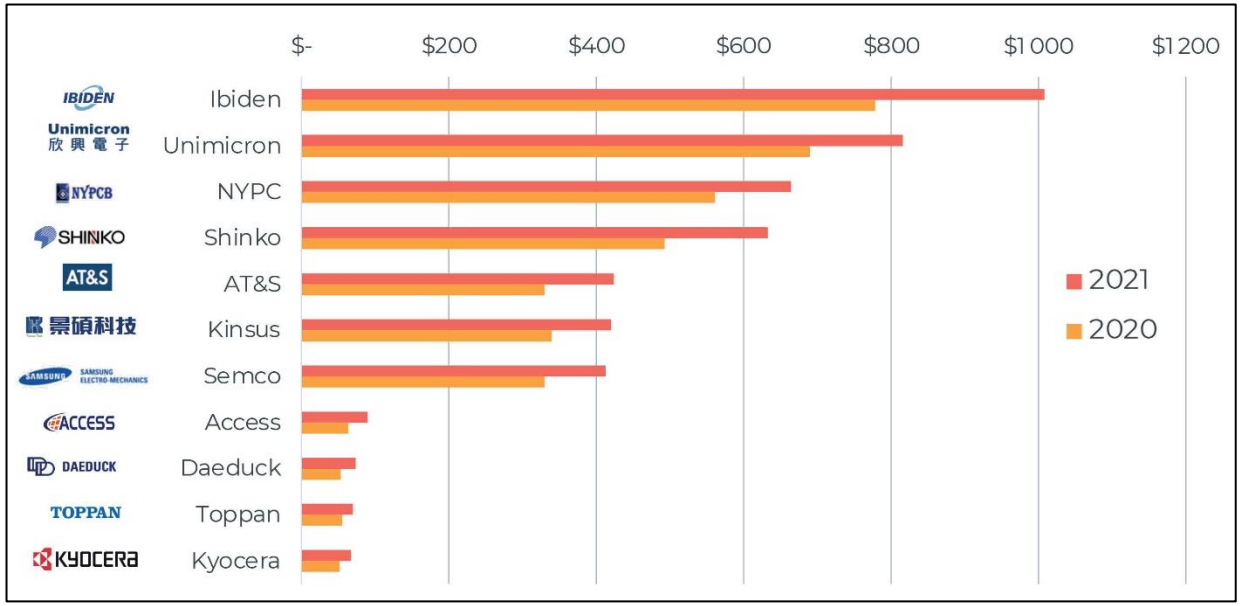


资料来源： KLA、中航证券研究所

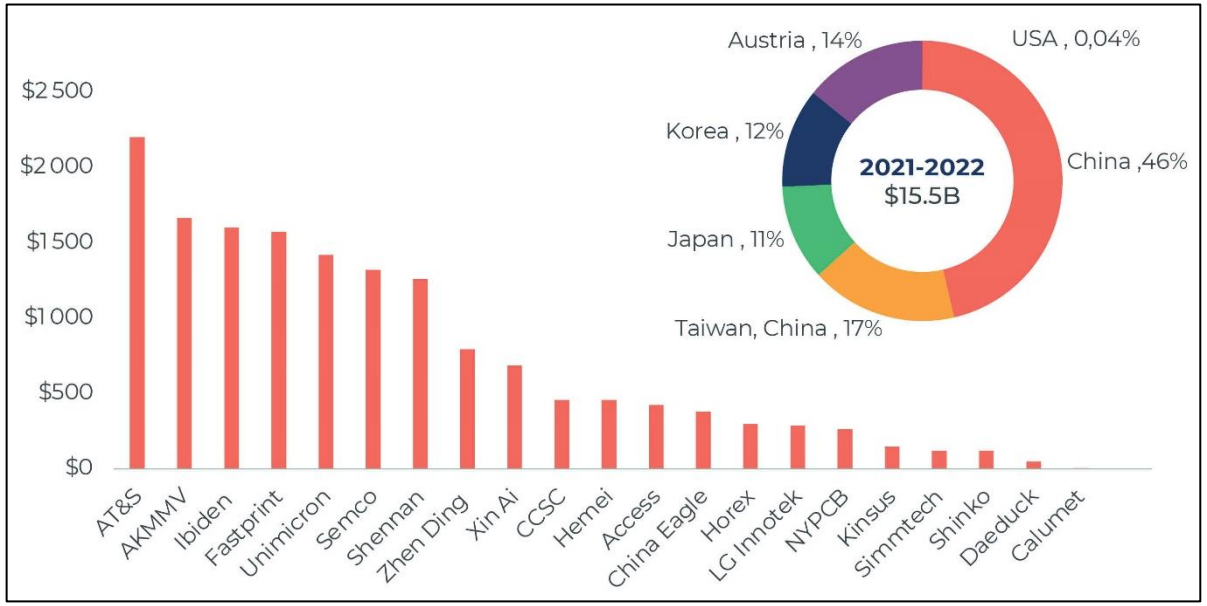
建议关注：兴森科技——引领IC载板国产替代

- 兴森科技专注于印制电路板产业，围绕传统PCB业务和半导体业务两大主线开展。PCB业务聚焦于样板快件及批量板的研发、设计、生产、销售和表面贴装；半导体业务聚焦于IC封装基板（含CSP封装基板和FCBGA封装基板）及半导体测试板。
- **兴森科技对IC载板产能投资全球领先。**据Yole，IC载板全球市场规模2022年达到151亿美元，2028年将增长至289亿美元，年均增速11%。但市场主要由海外公司主导，国内正在积极扩充产能,大陆厂商对IC载板的投资额占比达全球的46%，兴森科技（fastprint）的资本投入在2021-2022年间排名全球第四。

图：IC载板头部公司收入（百万美元）



图：中国IC载板产能投资领先全球（百万美元）



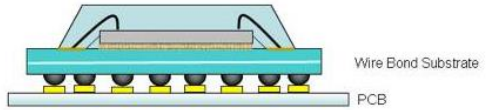
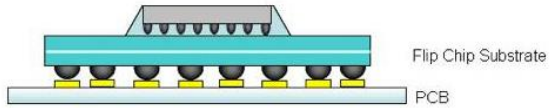
资料来源： yole、中航证券研究所

建议关注：兴森科技——引领IC载板国产替代

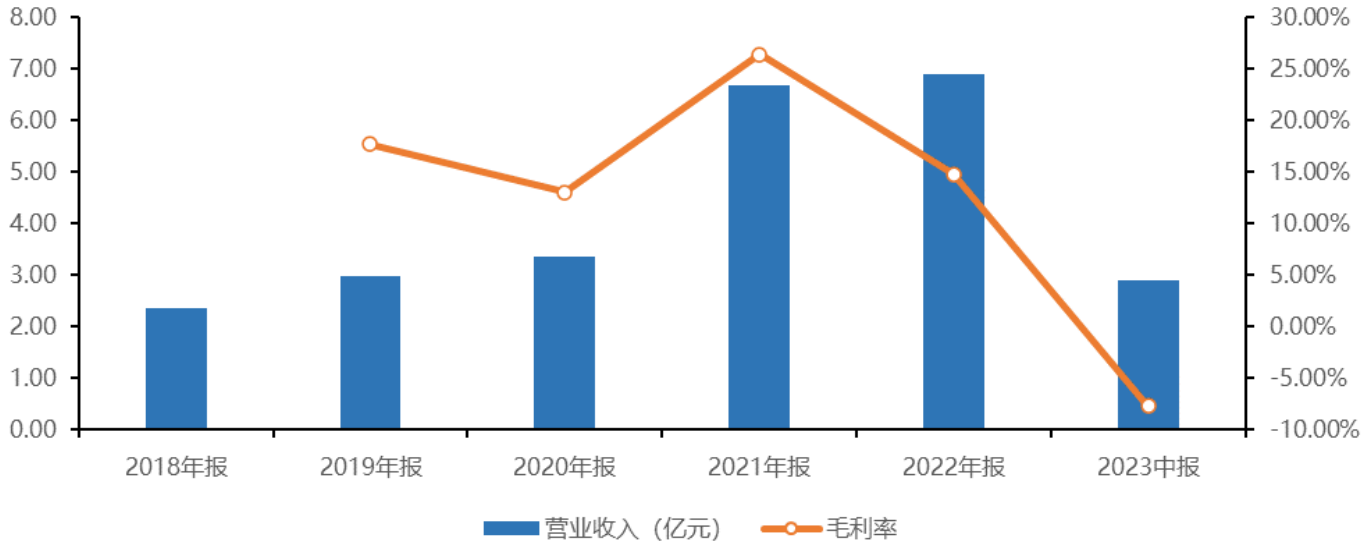
- 应用于智强手机的消费电子IC封装主要思索问题便携性、低成本等因素，普通认为合适而使用CSP封装，封装尺寸较小。应用于高性能计算机的IC封装，主要思索问题性能，普通认为合适而使用较为大型的、I/O数太多的BGA封装。

- 公司IC封装基板业务毛利率下降主要因行业需求大幅下滑导致整体产能利用率下降，随着行业回暖以及公司高端产能释放，IC载板业务有望好转。兴森珠海FCBGA封装基板项目拟建设产能200万颗/月（约6,000平方米/月）的产线，预计2024年第一季度进入小批量生产阶段。；广州FCBGA封装基板项目拟分期建设2,000万颗/月（2万平方米/月）的产线，一期厂房已于2022年9月完成厂房封顶；CSP封装基板产能为3.5万平方米/月，得益于行业需求逐步回暖，CSP封装基板产能利用率逐季提升。

图：主流的IC载板类型

IC载板	结构	应用
CSP（WB-CSP）		DRAM
PBGA（WB-BGA）		MCU、DSP
FC-CSP		Baseband、AP
FC-BGA		CPU、GPU、Chipset、ASIC

图：兴森科技IC载板收入及毛利率

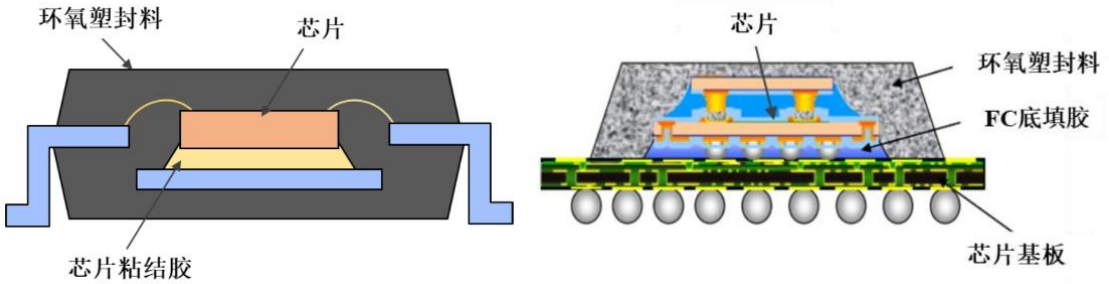


资料来源： 爱彼电路、ifind、中航证券研究所

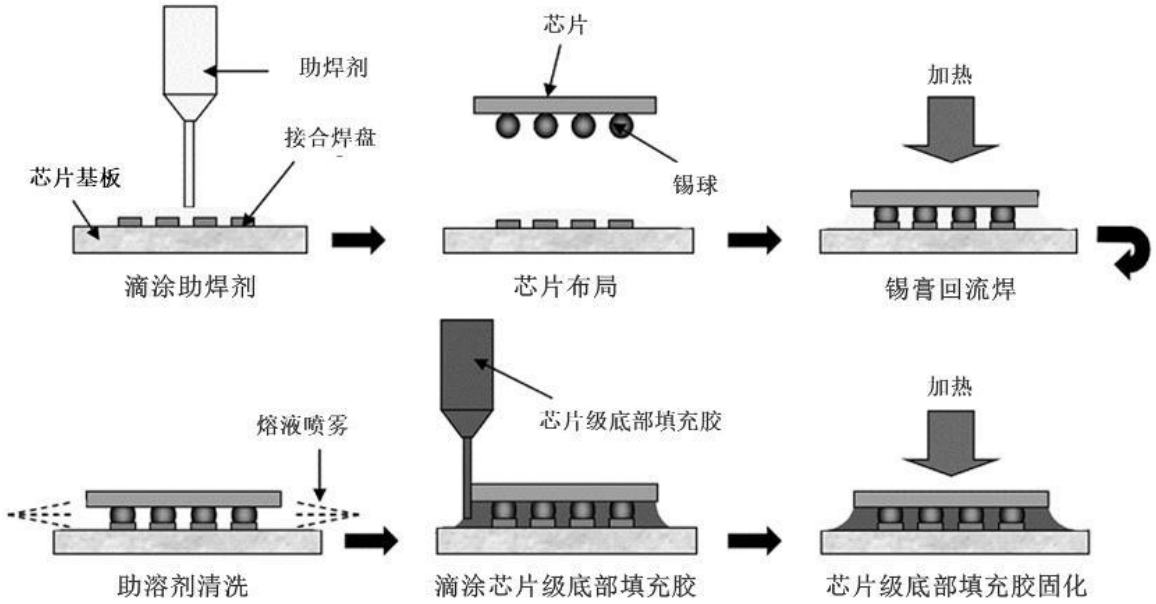
建议关注：华海诚科——环氧塑封领先企业

- 华海诚科主要产品包括环氧塑封料与电子胶黏剂，广泛应用于半导体封装、板级组装等应用场景。其中，环氧塑封料与芯片级电子胶黏剂与半导体封装技术的发展息息相关，是保证芯片功能稳定实现的关键材料。现建有先进的环氧模塑料中试线1条、大生产线5条。目前公司的研发能力和生产能力在国内环氧塑封料行业排名前列。
- 芯片级底部填充胶主要应用于FC（Flip Chip）封装领域，根据Yole，FC在先进封装的市场占比约为80%左右，是目前最具代表性的先进封装技术之一，具体类型包括FC-BGA、FC-SiP等先进封装技术，目前该市场仍主要为日本纳美仕、日立化成等外资厂商垄断，国内芯片级底部填充胶目前主要尚处于实验室阶段。公司FC底填胶已通过星科金朋的考核验证，在内资厂商中处于领先水平。

图：环氧塑封料应用场景



图：FC底填胶的使用流程

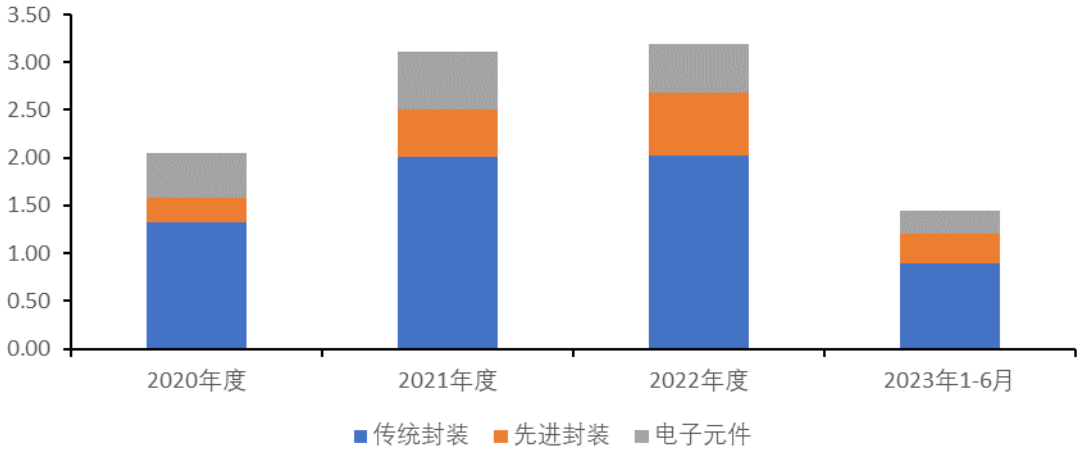


资料来源： 华海诚科招股书、中航证券研究所

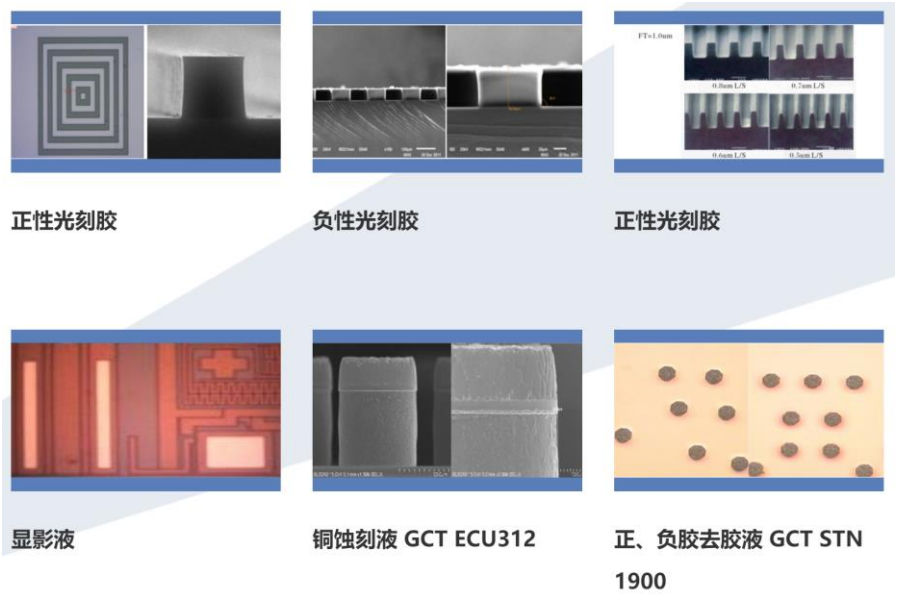
建议关注：艾森股份——先进封装材料平台

- 艾森股份围绕电子电镀、光刻两个半导体制造及封装过程中的关键工艺环节，形成了电镀液及配套试剂、光刻胶及配套试剂两大产品板块布局。公司主要产品以传统封装材料为主，先进封装材料逐步放量。公司先进封装用电镀铜基液（高纯硫酸铜）已在华天科技正式供应；先进封装用电镀锡银添加剂已通过长电科技的认证，尚待终端客户认证通过；先进封装用电镀铜添加剂正处于研发及认证阶段。先进封装光刻方面，公司以光刻胶配套试剂为切入点，成功实现附着力促进剂、显影液、去除剂、蚀刻液等产品在下游封装厂商的规模化供应。同时，公司积极开展光刻胶的研发，目前，公司自研先进封装用g/i线负性光刻胶已通过长电科技、华天科技认证并实现批量供应。

图：公司先进封装材料收入占比提升（亿元）



图：公司先进封装材料



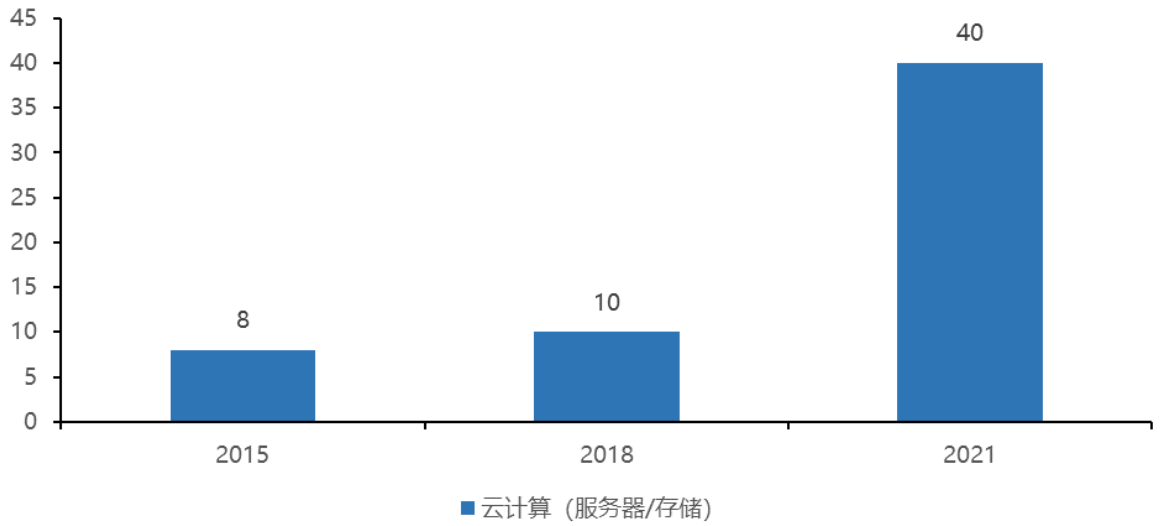
资料来源： 艾森股份招股书、中航证券研究所

目录

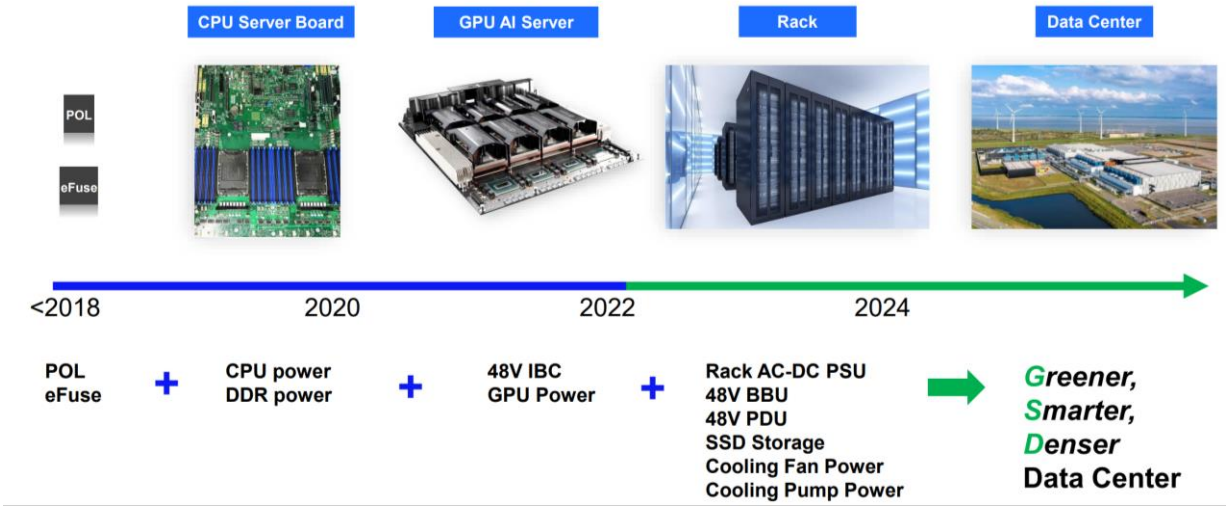
AI算力对高效电源提出新需求

- 据高性能电源解决方案供应商MPS测算，云计算领域的电源市场规模增长迅速，从2015年8亿美元，至2021年增长四倍，达到40亿美元。
- 云计算模拟电源市场由CPU服务器电路、GPU服务器电路、机架电路构成，市场规模分别为10亿美元、10亿美元、20亿美元。CPU服务器中，CPU供电、存储器供电、PoL（负载点）供电、eFuse（电子保险丝）市场规模分别为6亿、2.8亿美元、0.6亿美元、0.6亿美元。

图：图24 云计算领域的电源市场规模（亿美元）



图：云计算电源市场构成



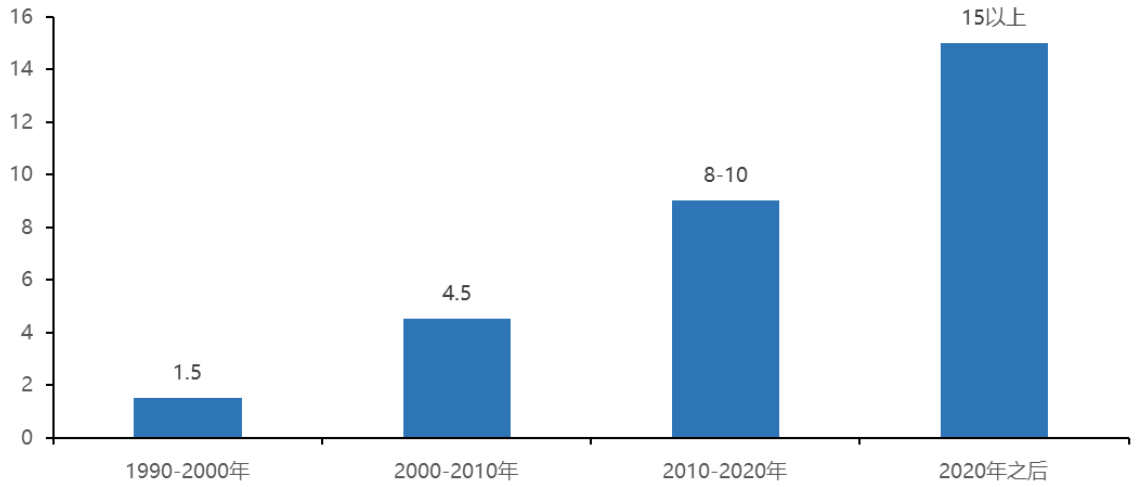
资料来源：MPS、中航证券研究所

服务器电源架构转向48V

- **数据中心功率密度增长，功耗随之提升。**随着服务器计算能力增长快速，总体功耗也随之提升。据华为，单IT机柜主流功率密度将从6-8kW/柜提高到12-15kW/柜@2027，超算、智算中心功率密度将提升至30kW以上。高达 40% 的数据中心运营成本来自为服务器机架供电与冷却所需的能源，由于数据中心功率密度增加，提高供电效率成为数据中心厂商创造额外利润的重要路径。

- **48V直流供电能够降低服务器配电损耗。**据松下测试结果，计算配电路径的电阻为0.1mΩ时的配电损耗，12V时为100W，48V时为6.25W，会出现16倍的计算差异。在各家公司都在积极致力于节省电力消耗中，谷歌公司于2016年率先引入了48V直流供电的手法。

图：中国数据中心功率密度趋势（kW）



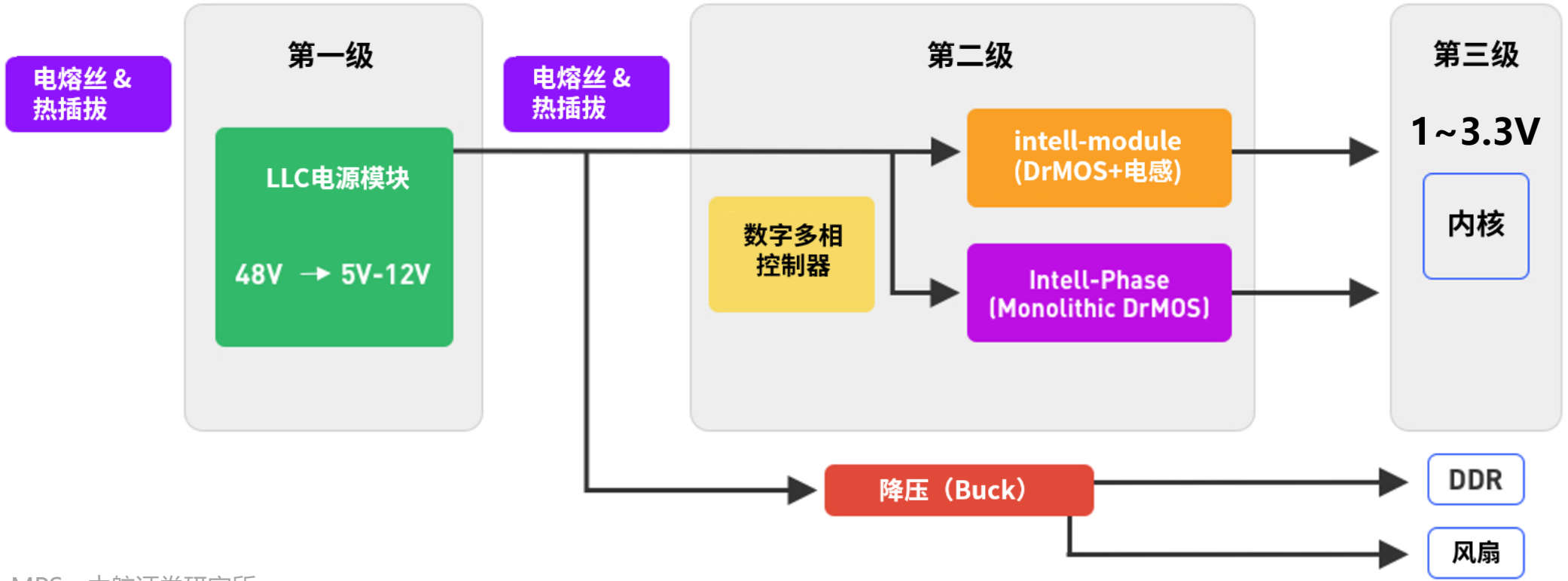
图：供电电压差异引起的损耗比较

资料来源：松下、中国信通院、中航证券研究所

算力电路的供电设计拥有较高壁垒

- 典型48V电源架构包含一个 AC/DC 前端，负责生成 48V 直流电。该直流电会被输送至DC/DC 变换器，在这里电压被转换成 12V 中间母线结构。然后将 12V 母线电压分配至板上多个负载点（PoL）变换器中，为内核CPU、GPU等提供电源。
- 大多数内核芯片或子电路所需的电压范围为 1V-3.3V，电流范围却为几十毫安至数百安培。通常，处理从几十毫安到数百安培应用程序的传统方式会采用离散模拟解决方案。离散模拟解决方案的构建块由一个控制器 IC 和一对外部 MOSFET 或一个驱动 MOSFET（DrMOS）IC 组成。由于内核芯片的电源轨对时序、电压精度、裕量和检测能力的设计要求非常严格，因此存在较大的设计壁垒。

图：服务器供电过程

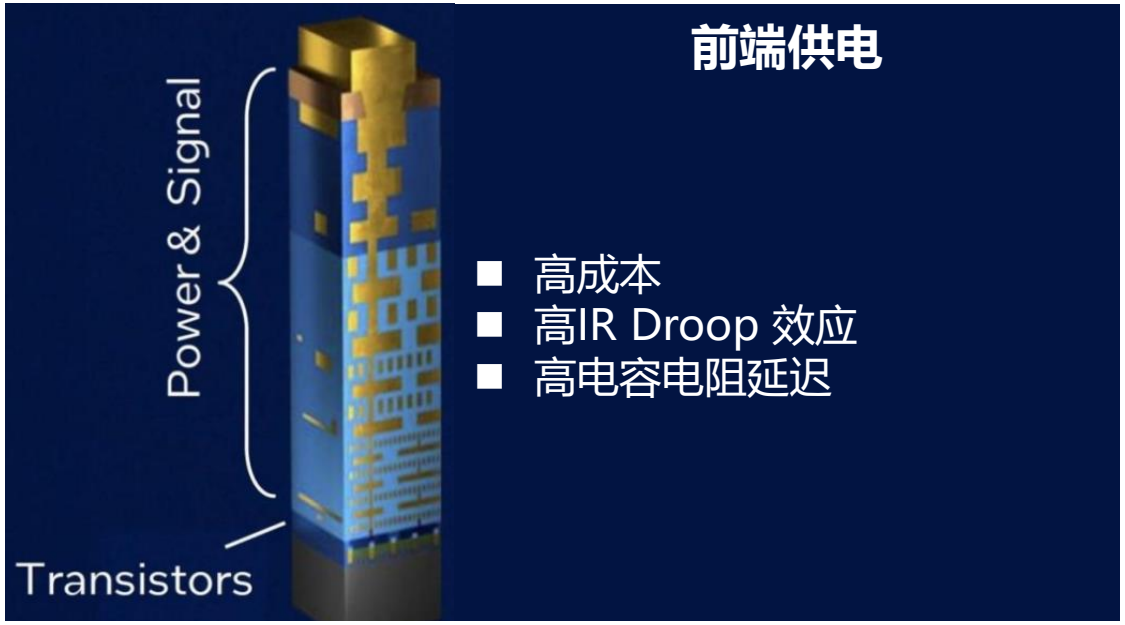


资料来源：MPS、中航证券研究所

背面供电技术提升堆叠芯片能效

- 随着晶体管越来越小，密度越来越高，互连线和电源线共存的线路层变成了一个越来越混乱的网络，堆栈层数也越来越多，可能需要穿过10至20层堆栈才能为下方的晶体管提供供电和数据信号。因此，业界在研究一种“将电源线迁移到芯片背面”的方法，这样可以让芯片正面只需要专注于与晶体管的信号互连。
- 背面供电技术（BSPDN）可解释成小芯片设计演变，原本将逻辑电路和存储器模组整合的现有方案，改成正面具备逻辑运算功能，背面供电或信号传递。背面供电技术能减轻线路后端的布线拥塞并提供电源性能优势，以及解决晶体管缩放中日益严重的电力输送问题。

图：前端供电 vs 背面供电

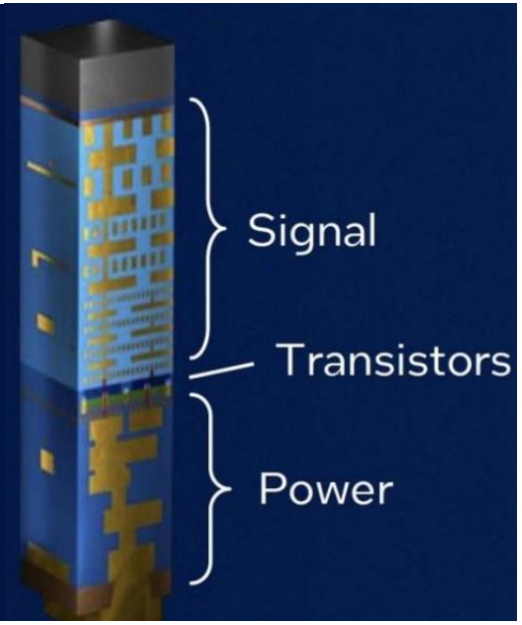


前端供电

- 高成本
- 高IR Droop 效应
- 高电容电阻延迟

背面供电

- 减少EUV光刻次数，降低成本
- IR Droop效应显著下降。IR降从50mV下降到20mV
- 芯片性能提高了44%，能效提高了30

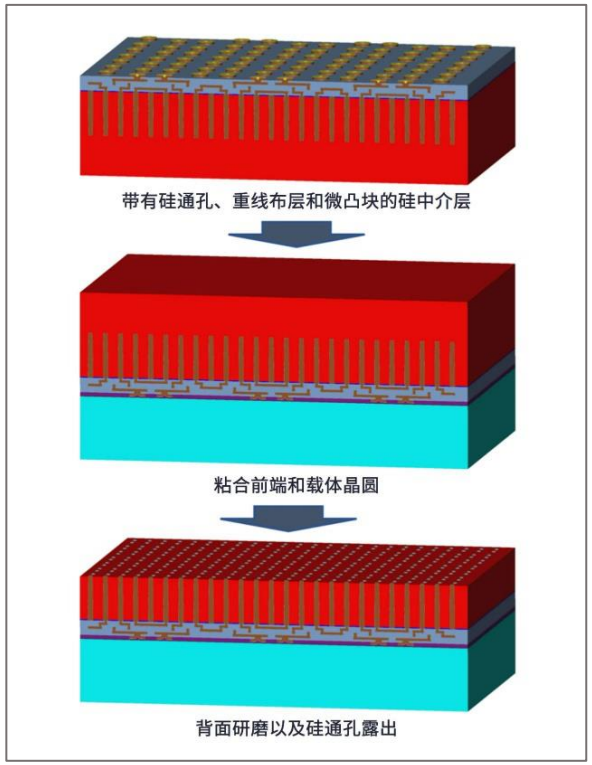


资料来源：半导体行业观察、维科网电子工程、中航证券研究所

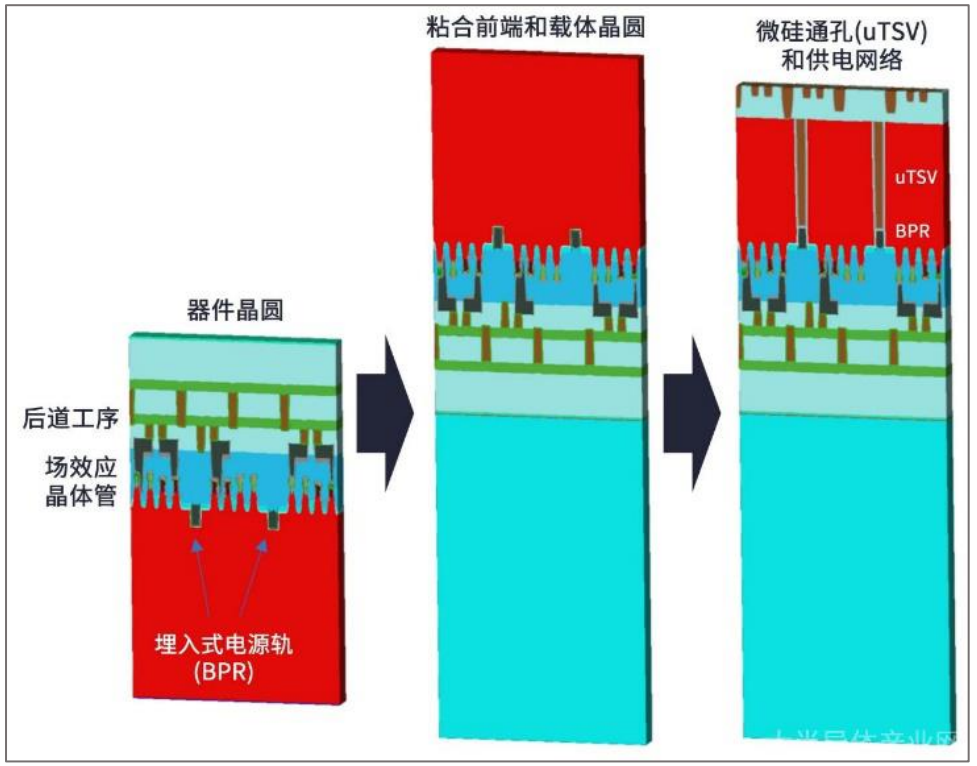
埋入式电源轨、TSV技术驱动背面供电发展

- 背部供电需要两项关键技术，分别是埋入式电源轨(BPR)与纳米硅穿孔(nTSV)。晶圆制造过程将先制造正面的晶体管，然后添加互联层，然后将晶圆反转，并对背面进行打磨减薄，再通过纳米硅穿孔（TSV）技术在晶圆背面进行制造供电网络，并与埋入式的电源轨连接。
- 埋入式电源轨是一种微缩化技术，可以进一步降低标准单元的高度，并减缓IR压降问题。这些电源轨是埋在电晶体下方的导线，一部份藏在硅基板内，另一部份则在浅沟槽隔离氧化层内。它们取代了传统后段制程在标准单元布下的电源线与接地线。

图：硅中介层的工艺处理



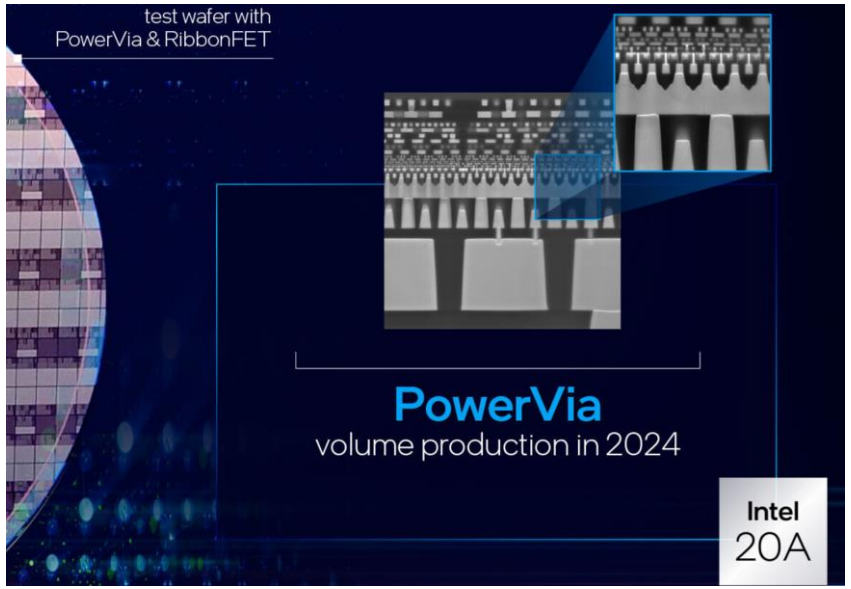
图：使用埋入式电源轨进行背面供电



资料来源：泛林集团、中航证券研究所

半导体巨头加码背面供电技术

- 台积电、三星、英特尔等芯片大厂都在积极布局背面供电网络技术，其中英特尔较为领先。英特尔将在Intel 20A工艺上首个采用PowerVia背面供电技术及RibbonFET全环绕栅极晶体管的节点，预计将于2024年上半年实现生产准备就绪，应用于未来量产的客户端ARL平台，目前正在晶圆厂启动步进。



Powervia背面供电技术预计将于2024年上半年实现生产准备就绪

2纳米背面电轨解决方案
计划于2025年下半年推出，
并在2026年实现量产

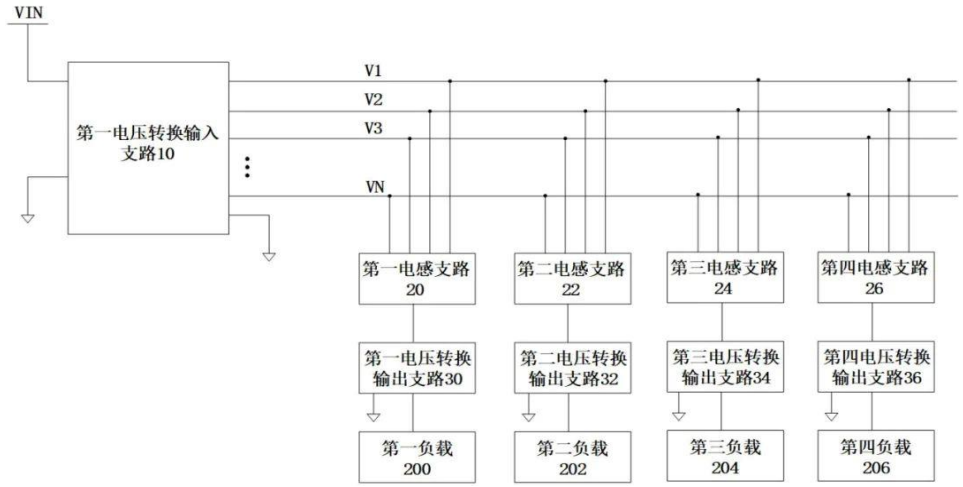
计划在2027年1.4nm工艺使用背面供电技术

资料来源：台积电、英特尔、三星、中航证券研究所

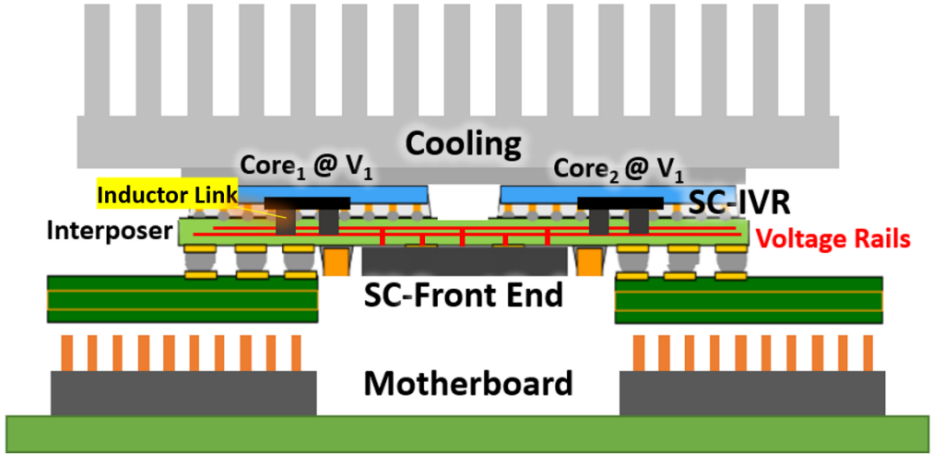
建议关注：希荻微——研发服务器高压供电方案

- 希荻微主要产品为服务于消费类电子和车载电子领域的电源管理芯片及信号链芯片等模拟集成电路，现有产品布局覆盖DC/DC芯片、锂电池充电管理芯片、端口保护和信号切换芯片、电源转换芯片等，具备高效率、高精度、高可靠性的良好性能。此外，公司计划拓展AF/OIS技术相关的音圈马达驱动芯片作为新的产品线。
- **希荻微与普林斯顿合作开发专利，解决服务器Chiplet供电瓶颈。**希荻微与普林斯顿两方的研究人员共同发明了一种基于混合型电荷泵电路的两级多输出供电架构。其中，电荷泵输入级以很高的效率将高压（如48V）输入转换为多个错相的脉冲，并通过多个电压轨和电感元件同时耦接到多个电荷泵输出级以为芯粒架构处理器的多个电压域分别供电。这样的架构显著地简化了供电电路，并为供电电路与处理器封装的进一步集成提供了一条新的路径。

图：希荻微与普林斯顿开发的专利架构图



图：采用希荻微专利技术的三维封装示意图



资料来源：希荻微官网、中航证券研究所

目 录

风险提示

- AI技术发展不及预期。AI算法、模型存较高不确定性，AI大模型迭代速度、升级效果存不确定性，AI技术整体进度发展可能延缓。
- AI应用不及预期。目前没有成熟的应用场景，使AI大模型形成好的商业闭环。
- 大模型成本过高的风险。目前AI大模型使用成本较高，如果下游接受不了长期高价的AI技术收费，相关AI投入增长或慢于预期，致使行业增长不及预期。
- 国产芯片发展不及预期的风险。目前国产GPU生态不成熟，设计和生产能力落后于国外，国内客户对国产芯片的接受程度有减弱的风险。
- 针对AI的监管政策收紧。AI可能带来越来越多的用户隐私、道德、伦理风险，由此将引致更严格的政策监管。

我们设定的上市公司投资评级如下：

买入 ：未来六个月的投资收益相对沪深300指数涨幅10%以上。
持有 ：未来六个月的投资收益相对沪深300指数涨幅-10%-10%之间
卖出 ：未来六个月的投资收益相对沪深300指数跌幅10%以上。

我们设定的行业投资评级如下：

增持 ：未来六个月行业增长水平高于同期沪深300指数。
中性 ：未来六个月行业增长水平与同期沪深300指数相若。
减持 ：未来六个月行业增长水平低于同期沪深300指数。

中航科技电子团队介绍：

首席：赵晓琨 SAC执业证书：S0640122030028
十六年消费电子及通讯行业工作经验，曾在华为、阿里巴巴、摩托罗拉、富士康等多家国际级头部品牌终端企业，负责过研发、工程、供应链采购等多岗位工作。曾任职华为终端半导体芯片采购总监，阿里巴巴人工智能实验室供应链采购总监。

分析师：刘牧野 SAC执业证书：S0640522040001
约翰霍普金斯大学机械系硕士，2022年1月加入中航证券。拥有高端制造、硬科技领域的投研经验，从事科技、电子行业研究。

研究助理 刘一楠 SAC执业证书：S0640122080006
西南财经大学金融硕士，2022年7月加入中航证券，覆盖半导体设备、半导体材料板块。

研究助理 苏弘宇 SAC执业证书：S0640122040021
俄亥俄州立大学金融数学学士，约翰霍普金斯大学金融学硕士。 2022年加入中航证券。

分析师承诺

负责本研究报告全部或部分内容的每一位证券分析师，再次申明，本报告清晰、准确地反映了分析师本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与，未来也将不会与本报告中的具体推荐或观点直接或间接相关。

风险提示：投资者自主作出投资决策并自行承担投资风险，任何形式的分享证券投资收益或者分担证券证券投资损失的书面或口头承诺均为无效。

免责声明