

端侧大模型近存计算，定制化存储研究框架

行业投资评级：强大于市|维持

中邮证券研究所 电子团队

分析师：吴文吉 S1340523050004

研究助理：翟一梦 S1340123040020

中邮证券

2025年2月19日

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

- **大模型赋能端侧AI。**在人工智能的飞速发展中，大型语言模型（LLMs）以其在自然语言处理（NLP）领域的革命性突破，引领着技术进步的新浪潮。自2017年Transformer架构的诞生以来，OpenAI的GPT系列到Meta的LLaMA系列等一系列模型崛起。这些模型传统上主要部署在云端服务器上，这种做法虽然保证了强大的计算力支持，却也带来了一系列挑战：网络延迟、数据安全、持续的联网要求等。这些问题在一定程度上限制了LLMs的广泛应用和用户的即时体验。正因如此，将LLMs部署在端侧设备上的探索应运而生，不仅能够提供更快响应速度，还能在保护用户隐私的同时，实现个性化的用户体验。端侧AI市场的全球规模正以惊人的速度增长，预计从2022年的152亿美元增长到2032年的1436亿美元，这一近十倍的增长不仅反映了市场对边缘AI解决方案的迫切需求，也预示着在制造、汽车、消费品等多个行业中，端侧AI技术将发挥越来越重要的作用。
- **存算一体技术的成熟为端侧AI大模型的商业化落地提供了技术基础。**作为一种新的计算架构，存算一体的核心是将存储与计算完全融合，存储器中叠加计算能力，以新的高效运算架构进行二维和三维矩阵计算，结合后摩尔时代先进封装、新型存储器件等技术，能有效克服冯·诺依曼架构瓶颈，实现计算能效的数量级提升。存算一体可分为近存计算(PNM)、存内处理(PIM)以及存内计算(CIM)。1) 近存计算通过将计算单元靠近内存单元，减少数据传输路径，提升访存带宽和效率，适合需要大规模并行处理和优化内存带宽的应用；2) 存内处理将计算单元嵌入存储芯片中，使存储器本身具备一定的计算能力，适合数据密集型任务，能够显著提升数据处理效率和能效比；3) 存内计算将存储单元和计算单元深度融合，使存储单元直接参与数据处理，适合高并行性计算和定制化硬件优化，能够消除数据访存延迟；在端侧AI大模型的商业化落地中，选择哪种技术取决于具体的应用需求和性能优化目标。
- **NPU赋能端侧大模型。**智能手机SoC自多年前就开始利用NPU(神经网络处理器)改善日常用户体验，赋能出色影像和音频，以及增强的连接和安全。不同之处在于，生成式AI用例需求在有着多样化要求和计算需求的垂直领域不断增加，这些AI用例面临两大共同的关键挑战：1) 在功耗和散热受限的终端上使用通用CPU和GPU服务平台的不同需求，难以满足这些AI用例严苛且多样化的计算需求；2) 这些AI用例在不断演进，在功能完全固定的硬件上部署这些用例不切实际。因此，**支持处理多样性的异构计算架构能够发挥每个处理器的优势**，例如以AI为中心定制设计的NPU，以及CPU和GPU。CPU擅长顺序控制和即时性，GPU适合并行数据流处理，NPU擅长标量、向量和张量数学运算，可用于核心AI工作负载。NPU降低部分易编程性以实现更高的峰值性能、能效和面积效率，从而运行机器学习所需的大量乘法、加法和其他运算。通过使用合适的处理器，异构计算能够实现最佳应用性能、能效和电池续航，赋能全新增强的生成式AI体验。
- **异构计算架构的实现需要先进封装技术的支持。**异构计算架构通过将不同功能的芯片（如CPU、GPU、FPGA、DSP等）或不同制程工艺的芯片集成在一起，实现高性能、高能效和多功能的计算系统，这种架构的实现需要先进的封装技术来支持。先进封装技术旨在通过创新的封装架构和工艺，提升芯片性能、降低功耗、减小尺寸，并优化成本。后文参考SiP与先进封装技术，将先进封装分为两大类梳理：①基于XY平面延伸的先进封装技术，主要通过**RDL**进行信号的延伸和互连；②基于Z轴延伸的先进封装技术，主要是通过**TSV**进行信号延伸和互连。

- **CUBE技术助力变革边缘AI计算。** 华邦电子开发的创新型CUBE (Customized Ultra Bandwidth Element, 定制化超高带宽元件) 技术, 作为客制化的高带宽存储芯片3D TSV DRAM, 专门为边缘AI运算装置所设计的存储架构, 利用3D堆叠技术并结合异质键合技术以提供高带宽、低功耗、单颗256Mb至8Gb的存储芯片, 并且可供模组制造商和SoC厂商直接部署。
- **CUBE架构:** CUBE是将SoC die置上(散热较好), DRAM die置下, 可以省去SoC中的TSV工艺, 进而降低了SoC die的尺寸与成本。同时, 3D DRAM TSV工艺可以将SoC信号引至外部, 使它们成为同一颗芯片, 进一步缩减了封装尺寸。
- **CUBE制造:** 由联电推动, 联电负责CMOS晶圆制造和晶圆对晶圆混合封装技术, 华邦电导入客制化CUBE架构, 智原提供全面的3D先进封装一站式服务, 以及存储IP和ASIC小芯片设计服务, 日月光则提供晶圆切割、封装和测试服务, 另外还有Cadence负责晶圆对晶圆设计流程, 提取TSV特性和签核认证。
- **CUBE容量及主要特性:**
 - ✓ 1) 基于D20工艺(20nm)的CUBE可以设计为1-8Gb/die 容量, 基于D16工艺的为16Gb/die 容量。非TSV和TSV堆叠均可用, 这为各种应用提供了优化内存带宽的灵活性。
 - ✓ 2) CUBE具有出色的能效, 在D20工艺中功耗低于1pJ/bit。
 - ✓ 3) CUBE的IO速度于1K I/O可高达2Gbps, 提供从16GB/s 至256GB/s 的总带宽。通过这种方式, CUBE能够确保带来高于行业标准的性能提升, 并通过uBump或混合键合增强电源和信号完整性。
 - ✓ 4) 基于D20标准的1-8Gb/die 产品, 以及灵活的设计和3D堆叠选择, 使得CUBE能够适应更小的外形尺寸。TSV的引入也进一步提高了性能, 改善了信号完整性、电源完整性和散热性能。TSV技术以及uBump/ 混合键合可降低功耗并节省SoC设计面积, 从而实现高效且极具成本效益的解决方案。利用TSV实现高效的3D堆叠, 简化了与先进封装技术的集成难度。通过减小芯片尺寸, CUBE能以更短的电源路径以及更紧凑、更轻巧的设计来降低器件成本、提高能效。
- **建议关注:**
 - ✓ 存储: 兆易创新
 - ✓ 数字: 瑞芯微, 寒武纪, 国科微, 北京君正, 全志科技, 炬芯科技
 - ✓ IP: 芯原股份
 - ✓ 封装: 长电科技, 通富微电, 华天科技, 甬矽电子, 晶方科技
- **风险提示:** AI端侧发展不及预期风险。



目录

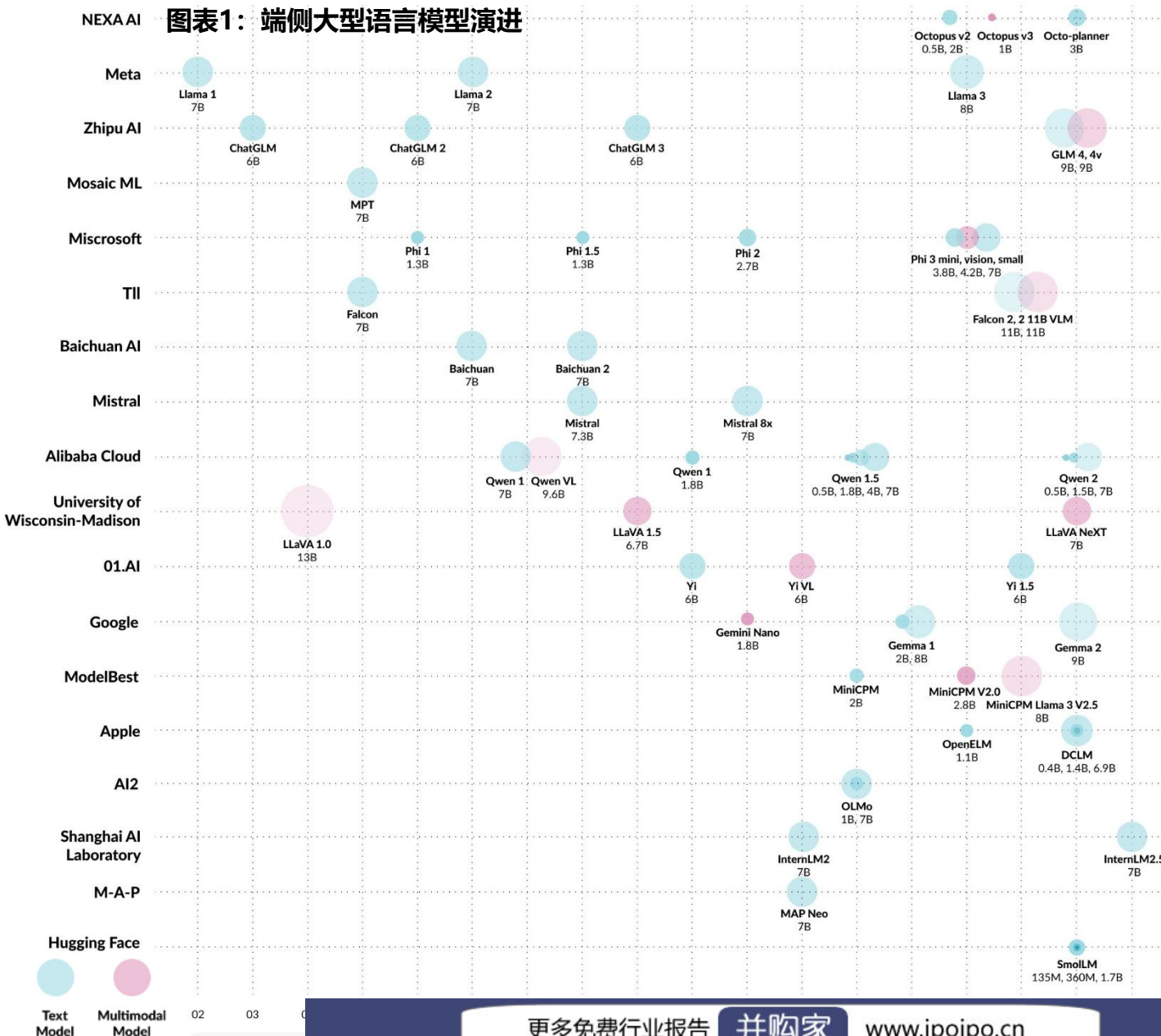
- | | |
|---|----------------|
| 一 | 端侧大模型 |
| 二 | 近存计算 |
| 三 | NPU |
| 四 | DRAM技术发展路径 |
| 五 | 先进封装 |
| 六 | 定制化存储：华邦CUBE介绍 |
| 七 | 相关标的 |



端侧大模型

端侧大型语言模型演进

图1：端侧大型语言模型演进



在人工智能的浪潮中，端侧大型语言模型（On-Device LLMs）迅猛发展且具备广泛的应用前景。自2023年起，随着参数量低于10B的模型系列如Meta的LLaMA、Microsoft的Phi系列等的涌现，LLMs在边缘设备上运行的可行性和重要性逐渐被验证。这些模型不仅在性能上取得了长足的进步，更通过混合专家、量化和压缩等技术，保持了参数量的优化，为边缘设备的多样化应用场景提供了强大支持。进入2024年，新模型的推出愈发密集，如左图所示，Nexa AI的Octopus系列、Google的Gemma系列等，不仅在文本处理上有所增强，更在多模态能力上展现了新的可能性，如结合文本与图像等多模态输入，以适应更复杂的用户交互需求。

资料来源：Jiajun Xu et al., "On-Device Language Models: A Comprehensive Review", 机器之心, 中邮证券研究所

请参阅附注免责声明

大语言模型架构基础

- **传统文本大型语言模型:** 从Transformer架构发展而来，最初由编码器和解码器组成。如今，流行的模型如GPT和LLaMA主要使用仅解码器架构。GPT模型在自注意力机制后应用层归一化，而LLaMA在每个子层前应用归一化以提高训练稳定性。在注意力机制方面，GPT模型使用标准自注意力机制，允许模型在生成序列时考虑输入序列中所有位置的信息，而LLaMA使用分组查询注意力(GQA)，优化计算和内存效率。混合专家(MoE)概念最早于1991年提出，在现代语言模型预训练中关键。MoE使用稀疏层减少计算资源，包含多个独立的“专家”网络和一个门控网络，以确定token的路由。
- **多模态大型语言模型:** 依托Transformer强大的学习能力，这些模型可以同时处理文本、图像、声音等多种模态。其内部运作机制如下：A) 使用标准交叉注意力层在模型内部层对多模态输入进行深度融合(如MultiModal-GPT)；B) 使用定制设计的层在模型内部层对多模态输入进行深度融合(LLaMA-Adapter, MoE-LLaVa)；C) 在模型输入阶段对多模态输入进行早期融合，使用特定模态的编码器(LLaVa, Qwen-VL)；D) 在输入阶段进行早期融合，但使用tokenization技术(如分词器)处理不同模态。
- **在资源有限的设备上部署大型语言模型面临内存和计算能力的挑战。**为解决这些问题，采用协作和分层模型方法分配计算负载。在资源受限设备上训练的经典方法包括量化感知缩放、稀疏更新、微型训练引擎(TTE)以及贡献分析。

图表2：端侧大语言模型训练的经典方法

方法	介绍
量化感知缩放	量化感知缩放:通过自动缩放不同位精度张量的梯度来稳定训练过程，解决量化图中不同位宽张量梯度尺度不一致的问题，使量化模型的训练精度与浮点模型相当
稀疏更新	选择性地更新网络中部分层的权重，跳过不太重要的层和子张量的梯度计算，从而减少内存使用和计算成本
微型训练引擎(TTE)	包括反向图中的冗余节点，如冻结权重的梯度节点，并重新排序操作以实现原位更新
贡献分析	自动确定稀疏更新方案，即确定哪些参数(权重/偏置)对下游精度贡献最大，以便在有限内存预算下选择应更新哪些层或张量部分

端侧大语言模型的性能指标

- 在评估设备端大型语言模型的性能时，有几个关键指标需要考虑：延迟、推理速度、内存使用以及存储和能耗，通过优化这些性能指标，设备端大型语言模型能够在更广泛的场景中高效运行，为用户提供更好的体验。同时硬件技术的持续进步显著影响了设备端大语言模型的部署和性能。

图表3：端侧大语言模型的性能指标

指标	介绍
延迟	是指从用户输入请求到系统开始响应所需的时间。通常使用TTFT（首次生成token时间）来衡量。延迟越低，用户体验越流畅。
推理速度	指模型基于已生成的所有token来预测下一个token的速度。由于每个新token都依赖于先前生成的token，因此推理速度对于用户对话的流畅性至关重要。
内存使用	使用的RAM/VRAM大小也是语言模型运行的性能指标之一。由于语言模型的运行机制，它们在推理过程中会根据模型参数的大小消耗相应的RAM。例如，在个人办公笔记本电脑上部署70B参数的模型是不切实际的。对于内存有限的设备，工程师需采用模型压缩技术来减少内存占用。
存储和能耗	模型占用的存储空间和推理过程中能耗对边缘设备尤为重要。在大多数情况下，大型语言模型推理会使处理器处于满负荷工作状态。如果运行时间过长，将严重消耗移动设备的电池。推理过程中的高能耗可能影响设备的电池寿命。例如，一个7B参数模型推理每个token将消耗约0.7J。对于电池容量约为50kJ的iPhone来说，这意味着与模型的对话最多只能持续两个小时。此外，模型推理引起的设备发热也是需要解决的问题。

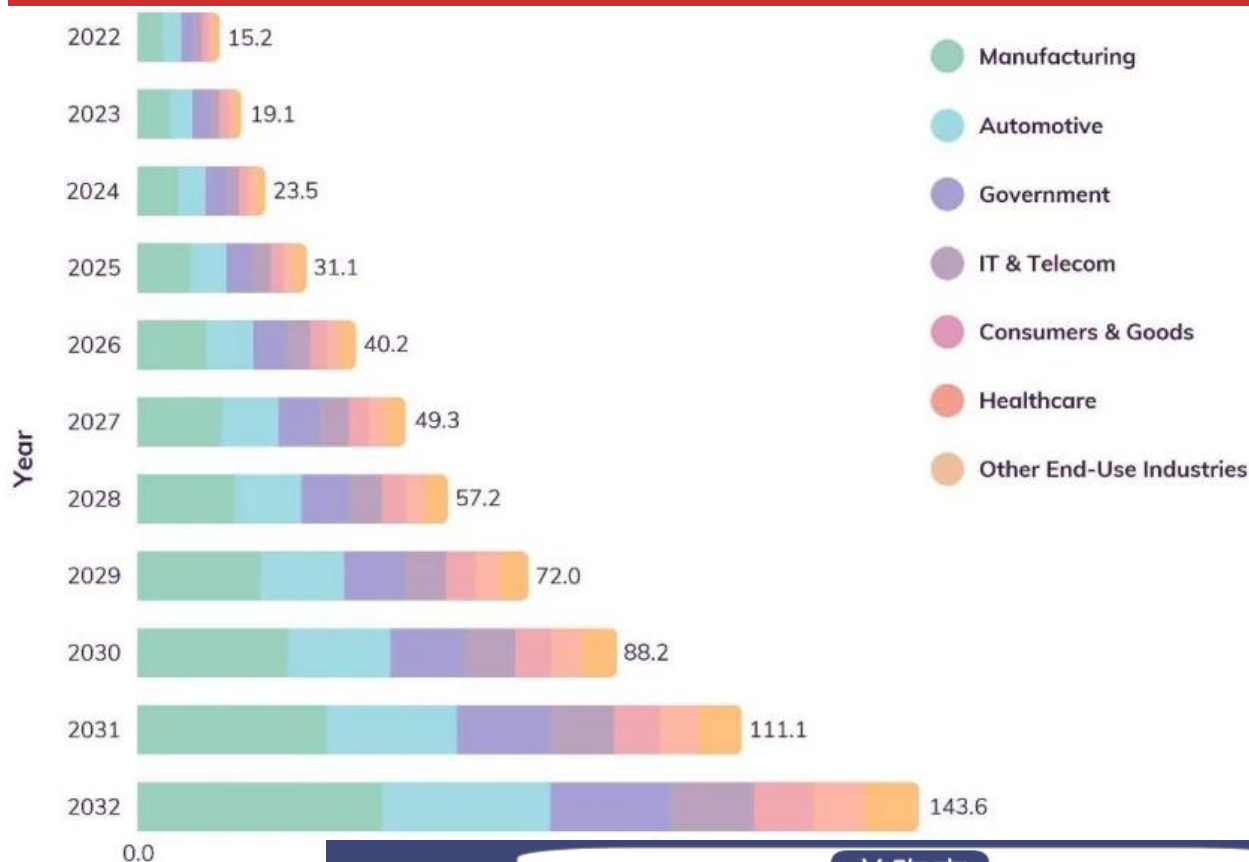
图表4：端侧大语言模型硬件介绍

硬件	介绍
GPU	凭借其大规模并行能力和高内存带宽，GPU已成为训练和加速大型语言模型的标准。NVIDIA的Tensor Cores在Volta架构中引入，并在后续几代中改进，为混合精度矩阵乘加运算提供了专门的硬件支持，这对基于Transformer的模型至关重要。NVIDIA的A100 GPU，配备80GB HBM2e内存，使得在单个设备上训练具有数十亿参数的模型成为可能。框架如Megatron-LM中实现的张量并行和流水线并行等技术，允许大语言模型在多个GPU上高效扩展。使用混合精度训练，特别是FP16和BF16格式，显著减少了内存占用，并增加了现代GPU上的计算吞吐量。
NPU	神经处理单元（NPU），也称为AI加速器，是专为机器学习工作负载设计的专用芯片。Google的张量处理单元（TPU）是一个突出的例子，v4版本每个芯片提供275 TFLOPS的BF16性能。TPU利用脉动阵列架构进行高效的矩阵乘法，特别适合大语言模型中的Transformer层。TPU Pod配置允许扩展到数千个芯片，使得训练如GPT-3和PaLM等大规模模型成为可能。其他NPU，如华为的昇腾AI处理器和Apple的Neural Engine，也通过量化和剪枝等技术为较小的大语言模型的设备端推理提供加速。
FPGA	现场可编程门阵列（FPGA）为加速大语言模型提供了灵活的硬件平台，尤其是在推理方面。最近的研究展示了在FPGA上高效实现Transformer层，利用稀疏矩阵乘法和量化等技术。例如，微软的Project Brainwave使用Intel Stratix 10 FPGA加速BERT推理，实现了低延迟和高吞吐量。FPGA在能效方面表现出色，可以针对特定模型架构进行优化，使其适合较小大语言模型的边缘部署。然而，与GPU和ASIC相比，FPGA的计算密度较低，限制了其在训练

边缘智能的新纪元

- 在人工智能的飞速发展中，大型语言模型（LLMs）以其在自然语言处理（NLP）领域的革命性突破，引领着技术进步的新浪潮。自2017年Transformer架构的诞生以来，OpenAI的GPT系列到Meta的LLaMA系列等一系列模型崛起。这些模型传统上主要部署在云端服务器上，这种做法虽然保证了强大的计算力支持，却也带来了一系列挑战：网络延迟、数据安全、持续的联网要求等。这些问题在一定程度上限制了LLMs的广泛应用和用户的即时体验。正因如此，将LLMs部署在端侧设备上的探索应运而生，不仅能够提供更快响应速度，还能在保护用户隐私的同时，实现个性化的用户体验。

图表5：2022年至2032年按终端用户划分的端侧AI全球市场规模（单位：十亿美元）

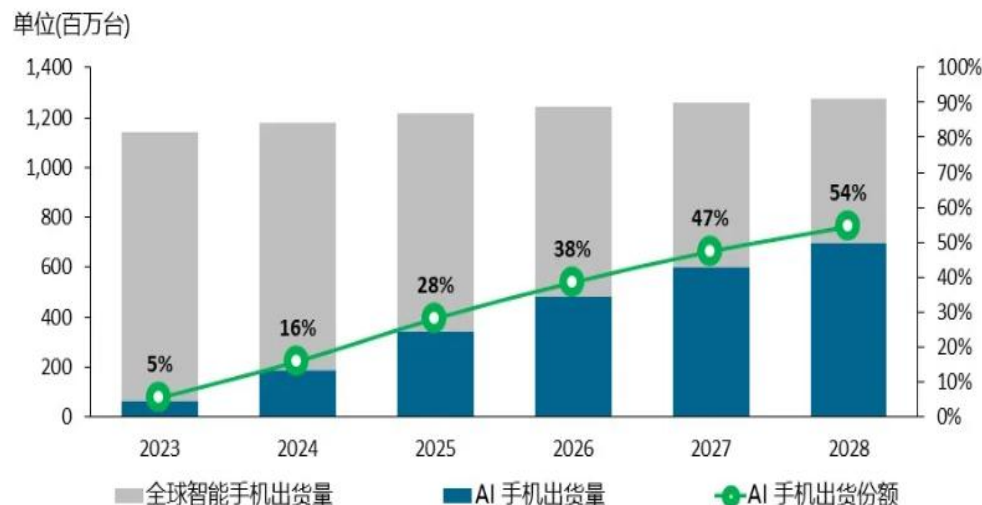


- 随着技术的不断进步，端侧AI市场的全球规模正以惊人的速度增长，预计从2022年的152亿美元增长到2032年的1436亿美元，这一近十倍的增长不仅反映了市场对边缘AI解决方案的迫切需求，也预示着在制造、汽车、消费品等多个行业中，端侧AI技术将发挥越来越重要的作用。

端侧AI出货量

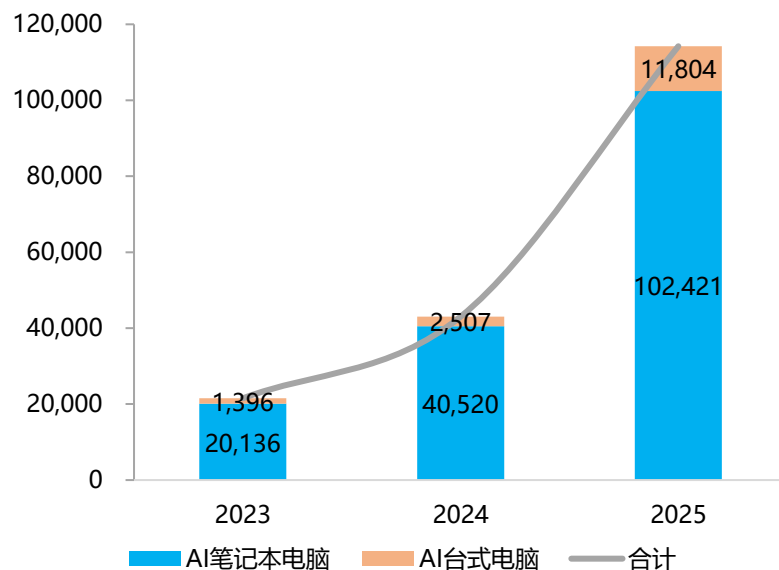
- **AI手机**: 在定义AI手机时,有几项核心硬件能力至关重要。对专用处理器,如ASIC、GPU以及其他零部件进行优化,以高效运行端侧AI模型和应用。根据Canalys预测,2024年,全球16%的智能手机出货为AI手机,到2028年,这一比例将激增至54%。受消费者对AI助手和端侧处理等增强功能需求的推动,2023年至2028年间,AI手机市场以63%的年均复合增长率(CAGR)增长。预计这一转变将先出现在高端机型上,然后逐渐为中端智能手机所采用,反映出端侧生成式AI作为更普适性的先进技术渗透整体手机市场的趋势。
- **AI PC**: Gartner将AI PC定义为带有嵌入式神经处理单元(NPU)的PC,并以此为基础进行预测。AI PC包括在ows on Arm、macOS on Arm和x86 on ows PC上安装NPU的PC。根据Gartner,2024年AI PC的出货量将达到4300万台,较2023年增长99.8%,2025年全球AI PC出货量将达到1.14亿台,较2024年增长165.5%,2025年,AI PC出货量在PC总出货量中的占比将从2024年的17%增长至43%;预计AI笔记本电脑的需求将高于AI台式电脑,2025年AI笔记本电脑的出货量将占到笔记本电脑总出货量的51%。

图表6: 2023-2028年全球AI手机出货量/份额预测数据



来源: Canalys (2024年5月)

图表7: 2023-2025年全球AI PC出货量 (千台)



来源: Gartner (2024年9月)

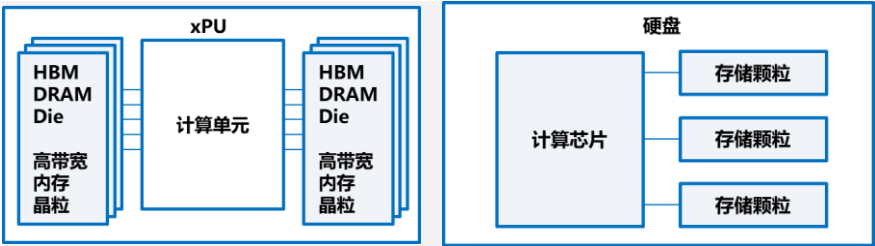
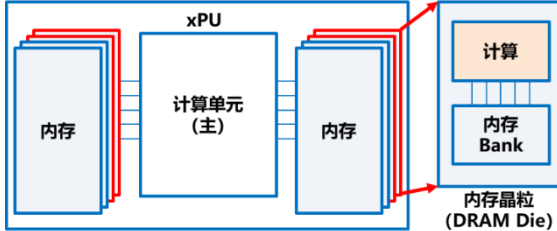


近存计算

存算一体技术分类

- 作为一种新的计算架构，存算一体的核心是将存储与计算完全融合，存储器中叠加计算能力，以新的高效运算架构进行二维和三维矩阵计算，结合后摩尔时代先进封装、新型存储器件等技术，能有效克服冯·诺依曼架构瓶颈，实现计算能效的数量级提升。存算一体可分为近存计算(PNM)、存内处理(PIM)以及存内计算(CIM)。

图表8：存算一体技术分类

存算一体技术分类	介绍	图示
近存计算 (PNM)	近存计算通过芯片封装和板卡组装等方式，将存储单元和计算单元集成，增加访存带宽、减少数据搬移，提升整体计算效率。近存计算仍是存算分离架构，本质上计算操作由位于存储外部、独立的计算单元完成其技术成熟度较高，主要包括存储上移、计算下移两种方式。近存计算已应用于人工智能、大数据、边缘计算等场景因其基本保持原有计算架构，产品化方案可较快投入使用。	 高带宽内存方案 可计算存储方案
存内处理 (PIM)	存内处理是在芯片制造的过程中，将存和算集成在同一个晶粒 (Die) 中，使存储器本身具备了一定算的能力。存内处理本质上仍是存算分离相比于近存计算，“存”与“算”距离更近。当前存内处理方案大多在内存 (DRAM) 芯片中实现部分数据处理，较为典型的产品形态为HBM-PIM和PIM-DIMM，在DRAM Die中内置处理单元，提供大吞吐低延迟片上处理能力，可应用于语音识别、数据库索引搜索、基因匹配等场景。	 基于DRAM的PIM方案实例
存内计算 (CIM)	存内计算即狭义的存算一体。在芯片设计过程中，不再区分存储单元和计算单元，真正实现存算融合。存内计算是计算新范式的研究热点，其本质是利用不同存储介质的物理特性，对存储电路进行重新设计使其同时具备计算和存储能力，直接消除“存”“算”界限，使计算能效达到数量级提升的目标。存内计算最典型的场景是为AI算法提供向量矩阵乘的算子加速，目前已经在神经网络领域开展大量研究，如卷积神经网络 (Convolutional Neural Network, CNN)、循环神经网络 (Recurrent Neural Network, RNN) 等。	在存储原位置上实现计算  CIM存内计算

近存计算

- 根据存储单元与计算单元融合的程度，可以分为近存计算和存内计算两类。虽然两类没有具体的界限，一个简单的分类如下：（1）近存计算设计，存储阵列一般无需改动，仍旧只提供数据的访存功能，而计算模块通常安放在存储阵列的附近；（2）存内计算设计，存储器件可以参与计算操作，这通常意味着存储阵列（memory cell array）需要改动来支持计算。可以按照存储器件工艺划分不同的技术路线，成熟存储工艺包括SRAM、DRAM、Flash等，新型存储工艺包括ReRAM、MRAM、PCRAM、FeRAM等。

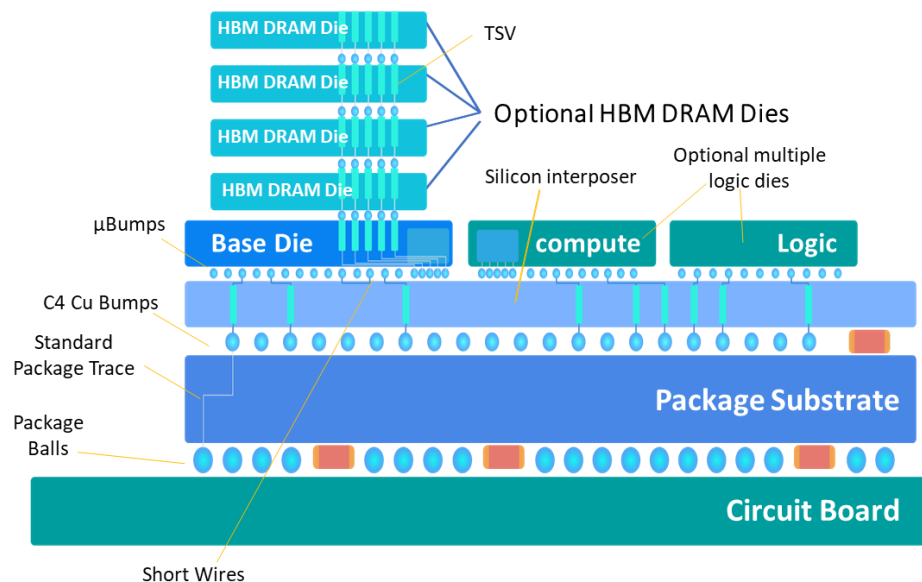
图表9：存算一体技术路径按照成熟存储工艺划分

SRAM	DRAM	Flash
优势是基于CMOS工艺，可以采用最先进的工艺节点，读写速度快；但是有存储密度低、静态漏电流高的缺点。基于SRAM的存算技术路线大致有三类： 基于SRAM的近存计算：通常指采用大量片上SRAM作为缓存的计算架构，计算采用数字方式、精度较高、通常面向大算力场景，代表：Graphcore、Tenstorrent等； 基于SRAM的数字存内计算：改造SRAM阵列，加入数字计算逻辑单元，在SRAM阵列中支持MAC计算，进一步提升Tensor计算的性能、减少功耗，适合AI大算力场景，代表：后摩智能、苹芯、TSMC等； 基于SRAM的模拟存内计算：改造SRAM宏单元，利用电流、电荷累计等模拟计算方式，支持MAC计算，在低精度计算场景有低功耗的优势，适合边缘/物联网等低算力、低功耗的场景。代表：九天睿芯；	优势是存储密度高于SRAM，适合数据中心等处理大容量模型的场景；但与CMOS工艺不兼容，访存性能和能效不如SRAM，其次设计需要DRAM vendor的支持。基于DRAM的存算技术路线大致有四类： 基于2D DRAM的近存计算：在DRAM芯片内部加入定制计算单元或者通用处理器，能够显著提升访存带宽，减少能耗，这种2D设计的好处是性价比高、可扩展性好，但是由于DRAM工艺的限制，能提供的计算密度受限，而且跨芯片间的通信带宽依旧受限，代表：Upmem、三星、海力士等； 基于2.5D DRAM的近存计算：利用2.5D集成技术，高性能计算芯片将HBM与处理单元集成在一起，提供大访存带宽，适用于大算力的场景，主要挑战是价格昂贵，功耗较高，代表：GPU、TPU、寒武纪等。 基于3D DRAM的近存计算：将计算单元与DRAM进行堆叠，甚至对HBM内部进行改造，把其中部分存储替换为计算单元，从而进一步提升带宽并减少访存功耗，相应的代价是增加了功耗密度、减少了存储容量等，代表：三星、平头哥等； 基于DRAM的存内计算：修改DRAM的存储阵列，来支持基本的计算逻辑，因为对DRAM修改较大，主要在学术界提出一些原型设计；	优势是存储密度高，但读写速度慢、擦写次数受限明显，基于Flash的存算技术路线大致有两类： 基于SSD的近存计算：也称为计算存储设备（Computational Storage Drive, CSD），在SSD控制器内/附近加入计算单元或者处理器，主要面向数据中心的大规模数据密集应用（如数据库，大数据分析等），代表：三星/Xilinx, ScaleFlux, NGD Systems等； 基于Flash的模拟存内计算：基于Flash的模拟存内计算功耗低，但是由于写入速度慢，且高精度（即每个cell存储多比特）数值写入有挑战，适合模型固定的低功耗应用场景，代表：知存科技、Mythic、闪忆科技等；

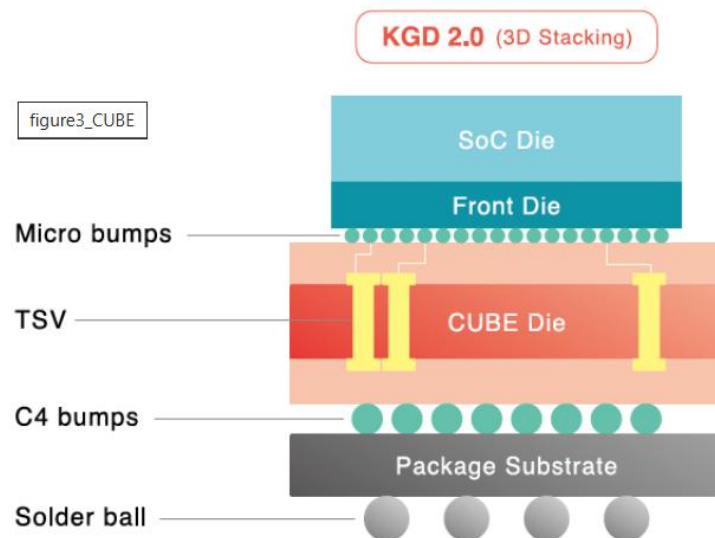
CUBE: 近存计算的一种方案，适合端侧AI

- 近存计算（PNM）主要包括存储上移、计算下移两种方式。存储上移指采用先进封装技术将存储器向处理器（如CPU、GPU）靠近，增加计算和存储间的链路数量，提供更高访存带宽。典型的产品形态为高带宽内存（High Bandwidth Memory, HBM），将内存颗粒通过硅通孔（Through Silicon Via, TSV）多层堆叠实现存储容量提升，同时基于硅中介板的高速接口与计算单元互联提供高带宽存储服务。计算下移指采用板卡集成技术将数据处理能力卸载到存储器，由近端处理器进行数据处理，有效减少存储器与远端处理器的数据搬移开销。典型的方案为可计算存储（Computational Storage Drives, CSD），通过在存储设备引入计算引擎承担如数据压缩、搜索、视频文件转码等本地处理，减少远端处理器（如CPU）的负载。
- 基于HBM和3D封装的AI芯片由于成本高、功耗高等因素，不适合端侧，CUBE作为一款高带宽、低功耗、紧凑尺寸、极具成本效益，以及可定制化的为近存计算解决方案，可供模组制造商和SoC厂商直接部署，可以满足端侧AI应用日益增长的需求。

图表10：基于HBM和3D封装的AI芯片 - 近存计算



图表11：CUBE 3D堆叠





NPU

GPU&GPGPU

- GPU，即图形处理器（Graphics Processing Unit），又称显示核心、视觉处理器、显示芯片，是一种专门在个人电脑、工作站、游戏机和一些移动设备（如平板电脑、智能手机等）上做图像和图形相关运算工作的微处理器。根据应用端，可将GPU分为移动端和桌面端，其中桌面端又分为服务器GPU和PC端GPU。从结构看，GPU通常包括图形显存控制器、压缩单元、BIOS、图形和计算整列、总线接口、电源管理单元、视频管理单元、显示界面。GPU的出现使计算机减少了对CPU的依赖，并解放了部分原本CPU的工作。最初，GPU负责渲染2D和3D图像、动画和视频，现已发展成为人工智能领域重要的核心硬件。

图表12：GPU行业发展历程

时间	类型	相关标准	代表产品	基本特征
80年代	图形显示	CGA, VGA	IBM S150	光栅生成器
80年代末	2D加速	GD1, DirectFB	86C911	2D图元加速
90年代初	部分3D加速	OpenGL (1.1-4.1), DirectX (6.0-11)	Glnt300sX	硬件T&L
90年代后期	固定管线		GeForce256	shader 功能固定
2004-2010年	统一渲染		G80	多功能 shader
2011年至今	通用计算	CUDA, OpenCL 1.2-2.0	TESLA	完成与图形处理无关的科学计算

- 对GPU通用计算进行深入研究从2003年开始，并出现了GPGPU概念，“GP”表示通用目的(General Purpose)，GPGPU一般也被称为通用图形处理器或通用GPU。NVIDIA于2007年率先推出了独立GPU（独显），使其作为“协处理器”在PC和服务端负责加速计算，承接CPU计算密集部分的工作负载，同时由CPU继续运行其余程序代码。
- 在GPU走向通用计算的过程中，统一渲染架构的出现非常关键。统一渲染单元是一个高性能的浮点和矢量计算逻辑，它具有通用和可编程属性。由此，GPU不再有单独的顶端渲染单元和像素渲染单元，而是由一个通用的渲染单元同时完成顶点和像素渲染任务。
- 基于统一渲染架构，GPU中的可编程计算单元Shader（着色器）core被挖掘出了更多的使用方法，比如通用计算。GPU从若干专用的固定功能单元（Fixed Function Unit）组成的专用并行处理器，进化为以通用计算资源为主、固定功能单元为辅的架构，这一架构的出现奠定了GPGPU的发展基础。

GPU&GPGPU

- 虽然都是由GPU架构演进而来，但关注的重点有明显区别，GPU的核心价值体现在图形图像渲染，GPGPU的重点在于算力。GPGPU架构设计时，去掉了为图形处理而设计的硬件加速单元，保留了GPU的SIMT架构和通用计算单元，使之更适合高性能并行计算，并能使用更高级别的编程语言，在性能、易用性和通用性上更加强大。从技术架构上看，GPU包含多个GPC，每个GPC包含多个TPC，TPC中包含多个SM，SM中包含CUDA核心和张量核心。GPGPU在GPU的基础上，增加专用向量、张量及矩阵运算指令，强化浮点运算的精度和性能。
- 随着人工智能大模型的快速发展，算力需求呈现出爆发式增长，传统的CPU芯片已经无法满足算力增长的需求，异构加速卡成为当前大模型领域最常用的计算硬件。当前大模型主要是使用的加速卡从架构上可以分为GPGPU和NPU两大阵营，其中GPGPU以国际大厂NVIDIA为代表，而NPU以国内厂商寒武纪MLU、华为Ascend系列等加速卡为代表。

图表13：人工智能基础硬件：GPGPU和NPU

GPGPU	NPU
GPGPU源自于图形计算领域，其发展较早，具有如下特点： ➢ GPGPU具备强大的并行计算能力，特别适合于处理AI计算中大量的矩阵运算任务。例如，在Transformer模型的训练和推理过程中，GPGPU能够显著加速计算过程，满足模型对高性能计算的需求； ➢ 由于发展较早GPGPU在深度学习领域拥有成熟的软件生态和编译工具链。这些工具和库为研究人员和工程师提供了丰富的API和框架，极大地方便了模型的开发和优化； ➢ GPGPU具有广泛的泛用性，其不仅适用于AI领域，还可以用于其他计算密集型任务，如大规模数据集的科学和工程计算等，这使得GPGPU具有更广泛的应用前景和市场需求。	NPU最初是专为深度学习和人工智能任务设计的专用处理器。与GPGPU不同，NPU在设计之初便专注于加速神经网络的推理和训练过程，其架构特点如下： ➢ 专门为加速神经网络计算而设计的芯片，能够高效地处理AI计算中的大量神经网络推理和训练任务。NPU通过集成大量的乘加单元和加大片内缓存，减少了数据IO瓶颈； ➢ 由于专注于特定任务，NPU的功耗通常比GPGPU更低，而在特定深度学习任务上的性能表现可能更优； ➢ NPU通常配备更多的片上内存（On-Chip Memory），以减少数据传输的延迟，提高数据处理效率。

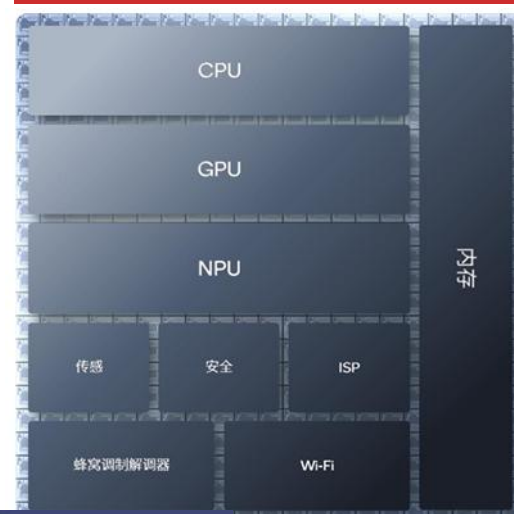
NPU：赋能端侧、边缘侧大模型

- 智能手机SoC自多年前就开始利用NPU(神经网络处理器)改善日常用户体验，赋能出色影像和音频，以及增强的连接和安全。不同之处在于，生成式AI用例需求在有着多样化要求和计算需求的垂直领域不断增加，这些用例可分为按需型用例、持续型用例以及泛在型用例。这些AI用例面临两大共同的关键挑战：1) 在功耗和散热受限的终端上使用通用CPU和GPU服务平台的不同需求，难以满足这些AI用例严苛且多样化的计算需求；2) 这些 AI用例在不断演进，在功能完全固定的硬件上部署这些用例不切实际。因此，支持处理多样性的异构计算架构能够发挥每个处理器的优势，例如以AI为中心定制设计的NPU，以及CPU和GPU。
- 每个处理器擅长不同的任务：CPU擅长顺序控制和即时性，GPU适合并行数据流处理，NPU擅长标量、向量和张量数学运算，可用于核心AI工作负载。CPU和GPU是通用处理器，为灵活性而设计，非常易于编程，“本职工作”是负责运行操作系统、游戏和其他应用等。而这些“本职工作”同时也会随时限制他们运行AI工作负载的可用容量。NPU专为AI打造，AI就是它的“本职工作”。NPU降低部分易编程性以实现更高的峰值性能、能效和面积效率，从而运行机器学习所需的大量乘法、加法和其他运算。通过使用合适的处理器，异构计算能够实现最佳应用性能、能效和电池续航，赋能全新增强的生成式 AI 体验。

图表14：生成式AI用例需求

生成式AI用例需求	介绍
按需型	按需型用例由用户触发，需要立即响应，包括照片/视频拍摄、图像生成/编辑、代码生成、录音转录/摘要和文本(电子邮件、文档等)创作/摘要。这包括用户用手机输入文字创作自定义图像、在 PC 上生成会议摘要，或在开车时用语音查询最近的加油站。
持续型	持续型用例运行时间较长，包括语音识别、游戏和视频的超级分辨率、视频通话的音频视频处理以及实时翻译。这包括用户在海外出差时使用手机作为实时对话翻译器，以及在 PC 上玩游戏时逐帧运行超级分辨率。
泛在型	泛在型用例在后台持续运行，包括始终开启的预测性AI助手、基于情境感知的 AI 个性化和高级文本自动填充。例如手机可以根据用户的对话内容自动建议与同事的会议、PC端的学习辅导助手则能够根据用户的答题情况实时调整学习资料。

图表15：SoC示意图



NPU：赋能端侧、边缘侧大模型

- NPU专为实现以低功耗加速AI推理而全新打造，并随着新AI用例、模型和需求的发展不断演进。对整体SoC系统设计、内存访问模式和其他处理器架构运行AI工作负载时的瓶颈进行的分析会深刻影响NPU设计。这些AI工作负载主要包括由标量、向量和张量数学组成的神经网络层计算，以及随后的非线性激活函数。
- 2015年，早期的NPU面向音频和语音AI用例而设计，这些用例基于简单卷积神经网络(CNN)并且主要需要标量和向量数学运算。从2016年开始，拍照和视频AI用例大受欢迎，出现了基于Transformer、循环神经网络(RNN)、长短期记忆网络(LSTM)和更高维度的卷积神经网络(CNN)等更复杂的全新模型。这些工作负载需要大量张量数学运算，因此NPU增加了张量加速器和卷积加速，让处理效率大幅提升。有了面向张量乘法的大共享内存配置和专用硬件，不仅能够显著提高性能，而且可以降低内存带宽占用和能耗。例如，一个 $N \times N$ 矩阵和另一个 $N \times N$ 矩阵相乘，需要读取 $2N^2$ 个值并进行 $2N^3$ 次运算(单个乘法和加法)。在张量加速器中，每次内存访问的计算操作比率为 $N:1$ ，而对于标量和向量加速器，这一比率要小得多。
- 2023年，大语言模型(LLM)--比如Llama 2-7B，和大视觉模型(LVM)--比如Stable Diffusion 赋能的生成式AI使得典型模型的大小提升超过了一个数量级。除计算需求之外，还需要重点考虑内存和系统设计，通过减少内存数据传输以提高性能和能效。
- 2024年，NPU在多模态大模型中的应用逐渐普及，支持更高效的推理和更自然的交互。随着AI持续快速演进，必须在性能、功耗、效率、可编程性和面积之间进行权衡取舍。一个专用的定制化设计NPU能够做出正确的选择，与AI行业方向保持高度一致。

图表16：NPU随着不断变化的AI用例和模型持续演进，实现高性能低功耗



异构计算：利用全部处理器支持生成式AI

- 正如前述，大多数生成式 AI 用例可分类为按需型、持续型或泛在型用例。按需型应用的关键性能指标是时延，因为用户不想等待。这些应用使用小模型时，CPU 通常是正确的选择。**当模型变大(比如数十亿参数)时，GPU 和 NPU 往往更合适。电池续航和能效对于持续和泛在型用例至关重要，因此 NPU 是最佳选择。**
- 另一个关键区别在于 AI 模型为内存限制型(即性能表现受限于内存带宽)，还是计算限制型(即性能表现受限于处理器性能)。当前的大语言模型在生成文本时受内存限制，因此需要关注 CPU、GPU 或 NPU 的内存效率。对于可能受计算或内存限制的大视觉模型，可使用 GPU 或 NPU，但 NPU 可提供最佳的能效。
- 提供自然语音用户界面(UI)以提高生产力并增强用户体验的个人助手预计将成为一类流行的生成式 AI 应用。**语音识别、大语言模型和语音模型必将以某种并行方式运行，因此理想的情况是在 NPU、GPU、CPU 和传感处理器之间分布处理模型。**对于 PC 来说，个人助手预计将始终开启且无处不在地运行，考虑到性能和能效，应当尽可能在 NPU 上运行。

图表17：正如在工具箱中选择合适的工具，选择合适的处理器取决于诸多因素



- 适合终端侧执行的生成式 AI 模型日益复杂，参数规模也在不断提升，从 10 亿参数到 100 亿，甚至 700 亿参数。其多模态趋势日益增强，这意味着模型能够接受多种输入形式--比如文本、语音或图像，并生成多种输出结果。此外，许多用例需要同时运行多个模型。例如，个人助手应用采用语音输入输出，这需要运行一个支持语音生成文本的自动语音识别(ASR)模型、一个支持文本生成文本的大语言模型、和一个作为语音输出的文本生成语音(TTS)模型。**生成式 AI 工作负载的复杂性、并发性和多样性需要利用 SoC 中所有处理器的能力。**最佳的解决方案要求：1) 跨处理器和处理器内核扩展生成式 AI 处理；2) 将生成式 AI 模型和用例映射至一个或多个处理器及内核。
- **选择合适的处理器取决于众多因素，包括用例、终端类型、终端层级、开发时间、关键性能指标(KPI)和开发者的技术专长。**制定决策需要在众多因素之间进行权衡，针对不同用例的 KPI 目标可能是功耗、性能、时延或可获取性。

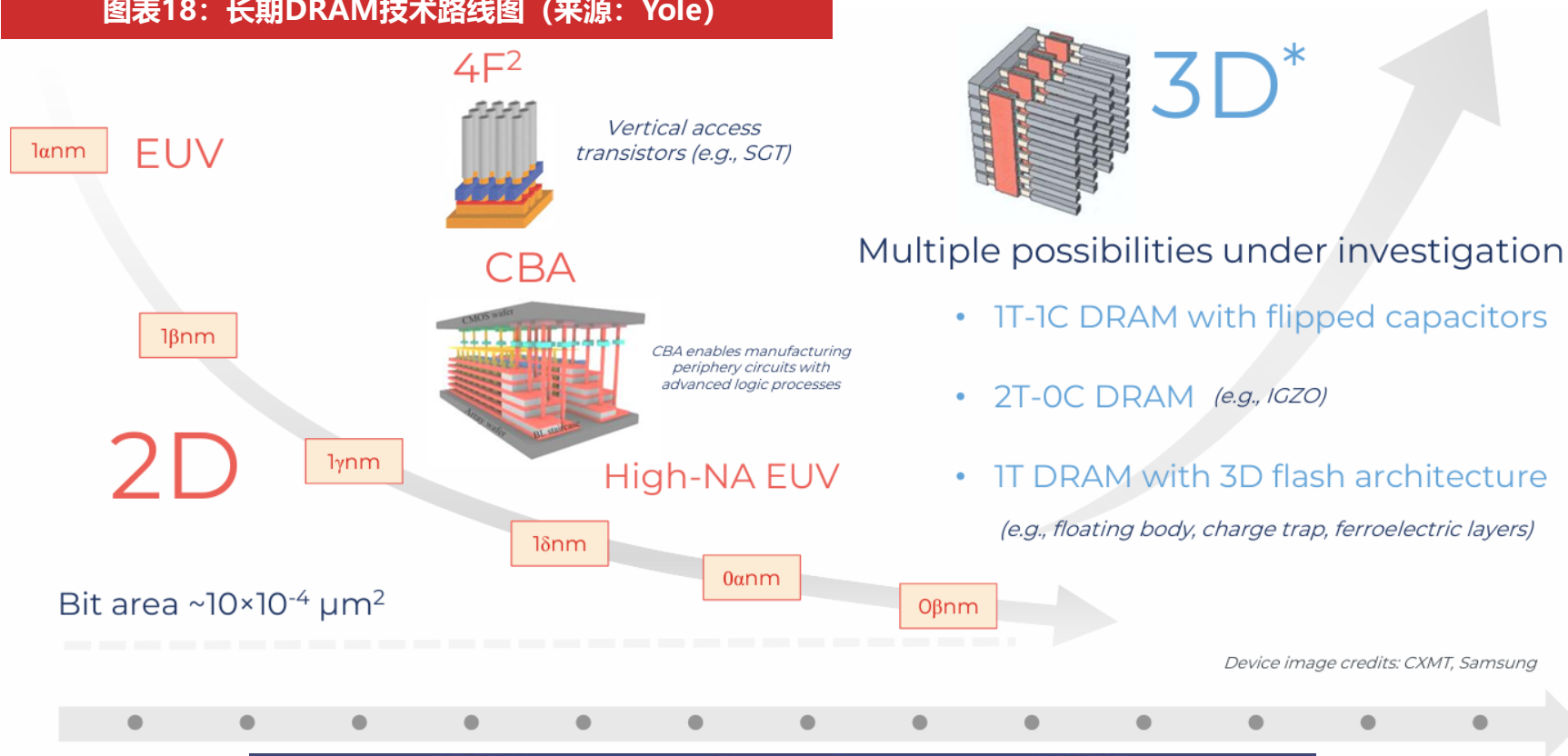
四

DRAM技术发展路径

长期DRAM技术发展

- **制程技术的持续微缩：**制程技术的微缩是DRAM发展的核心方向之一。目前，三星、SK海力士和美光等主要厂商已经进入1Znm（10-14nm）制程，并计划在未来几年内进一步缩小至1αnm（10nm以下）。2月18日，TechInsights发布最新报告，揭示了DRAM技术的未来发展趋势：到2027年底，DRAM预计将迈入个位数纳米技术节点，如D0a、D0b和D0c世代。这一突破将为AI和数据中心带来革命性的变革。
- **3D DRAM架构的兴起：**为了进一步提升存储密度，3D DRAM架构成为未来的关键发展方向。3D DRAM通过垂直堆叠存储单元，能够在不增加芯片面积的情况下显著提高存储密度。主要技术包括4F²垂直沟道晶体管（VCT）、IGZO DRAM单元和3D堆叠DRAM单元等，这些技术将在10nm以下级别实现产品化。

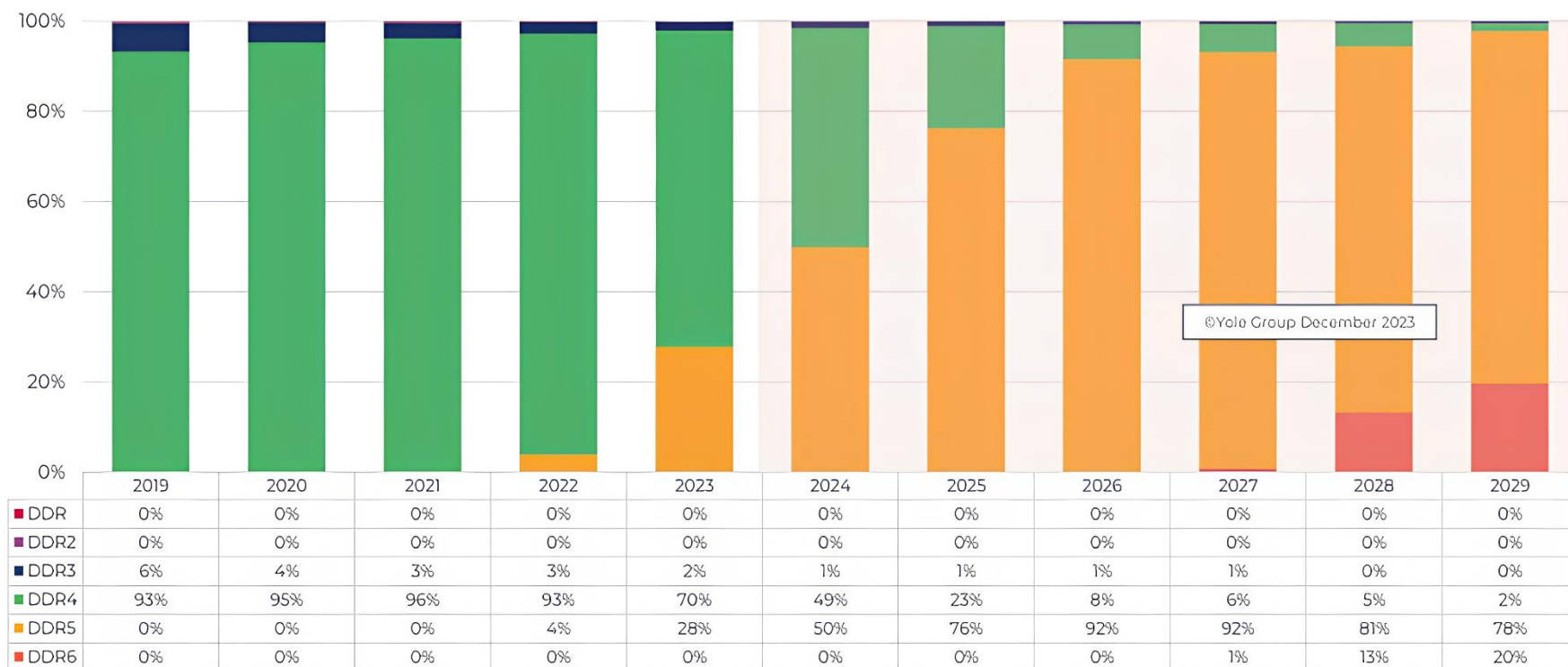
图表18：长期DRAM技术路线图（来源：Yole）



长期DRAM技术发展

- **高性能内存标准的演进：**DDR内存技术一直是主流的高性能内存标准，从DDR1到DDR5，每一代都显著提升了数据传输速率、降低了功耗，并优化了性能。作为最新一代标准，DDR5的传输速率从4800MT/s起步，相比DDR4的3200MT/s，带宽提升了50%。与DDR4相比，DDR5具有更高的速度、更大的容量和更低的功耗。此外，DDR5还集成了电源管理IC（PMIC），改善了信号完整性和功耗表现。随着AI和数据中心对内存带宽需求的增加，DDR6等下一代标准也在研发中，预计将进一步提升数据传输速率和容量。

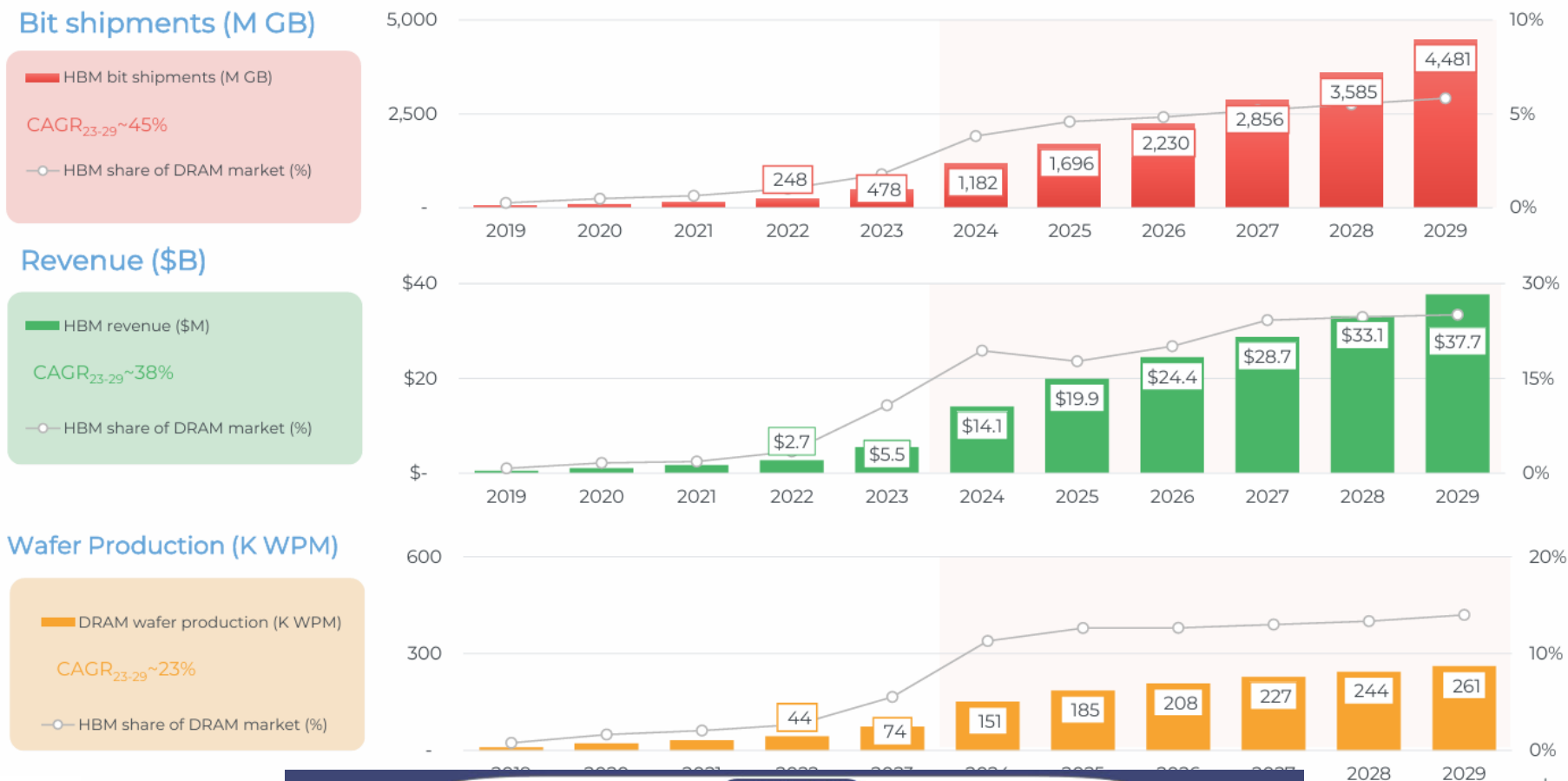
图表19：DDR DRAM bit 出货量（来源：Yole）



长期DRAM技术发展

- **高带宽内存 (HBM) 的发展:** HBM (高带宽内存) 技术通过将多个DRAM裸片堆叠并与GPU封装在一起,本质上是一种3D DRAM, 通过将多层DRAM芯片垂直堆叠, 并利用TSV技术实现层间互连, 实现了大容量、高位宽的内存组合。HBM技术在高性能计算、人工智能等领域具有显著优势。未来, HBM将集成更多创新特性, 例如采用新型材料 (如IGZO, 氧化铟镓锌) 和更先进的3D堆叠技术 (如垂直沟道晶体管VCT和无电容3D DRAM)。此外, HBM的未来发展还包括更高的堆叠层数、更低的功耗以及更广泛的应用场景。

图表20: HBM市场相关预测 (来源: Yole)



长期DRAM技术发展

- **新型材料与工艺的应用：**新型材料如IGZO（氧化铟镓锌）和相变材料的应用，将有助于降低DRAM的功耗并提升性能。先进制造工艺，如极紫外光刻（EUV）技术，也将成为实现更高密度DRAM的关键。

图表21：DRAM新型材料介绍

新型材料	介绍
IGZO（铟镓锌氧化物）	SK海力士正在研究将IGZO作为3D DRAM的新一代沟道材料。IGZO是一种金属氧化物材料，具有低待机功耗和稳定的物理、化学特性，适合长续航时间的DRAM应用。
高k介电材料	高k材料（如氧化铪）通过提高电介质的介电常数，可以在不增加电容器尺寸的情况下提升其电容值，从而缓解由于电介质厚度减少带来的电容下降问题。此外，高k金属栅极（HKMG）晶体管正在被引入到先进的DRAM中，以优化性能、功率、面积和成本。
低k介电材料（如Black Diamond™）	应用材料公司推出的Black Diamond™是一种低k介电材料，用于减少互连布线所需的晶粒面积，防止金属线之间的电容耦合，从而降低功率损耗、提高性能和可靠性。

图表22：DRAM新型工艺介绍

新型工艺	介绍
极紫外光刻（EUV）技术	EUV光刻技术通过使用极短波长的光束，能够在更小的尺度上进行精确的图案刻蚀，使得更高密度的晶体管布局成为可能。
垂直通道晶体管（VCT）	三星正在研究垂直通道晶体管技术，将传统水平方向的通道改为垂直方向，并采用栅极包裹通道的设计。这种设计可以显著减小器件面积，提高存储密度和性能。
3D堆叠技术	3D DRAM通过垂直堆叠存储单元，显著提高了存储密度和性能。例如，NEO半导体推出的3D X-DRAM技术，采用类似于3D NAND的单元阵列结构，通过多层掩模形成垂直堆叠的存储单元，实现了更高的存储密度和更低的制造成本。
先进刻蚀工艺	应用材料公司推出的Draco™硬掩模材料和Sym3™ Y刻蚀系统，通过协同优化提高了刻蚀选择比，使掩模更薄，并实现了更均匀的刻蚀效果。

长期DRAM技术发展

- **定制化与细分市场：**随着应用场景的多样化，DRAM市场将更加注重定制化解决方案。例如，服务器专用、汽车电子用和消费电子用DRAM将更加丰富多样。数据安全需求的提升也将推动具备内置加密功能的DRAM成为市场关注的焦点。

图表23：DRAM定制化发展案例

应用市场	举例
车规级DRAM市场	北京君正通过并购ISSI (Integrated Silicon Solution, Inc.) 进入车规级存储芯片市场。ISSI的DRAM产品涵盖从16M到16G等多种容量规格，能够满足工业、消费、通讯等级和车规等级的要求，具备在极端环境下稳定工作以及节能降耗的特点。这种定制化的产品策略使其在汽车电子领域获得了显著的市场份额。
数据中心与服务器DRAM	随着数据中心对带宽和性能的需求增长，DRAM厂商正在开发定制化的高性能内存解决方案。例如，SK海力士推出了MCR-DIMM（多路复用器组合等级双列直插式内存模块），该技术允许高端服务器DIMM以最低8Gbps的数据速率运行，相比现有DDR5内存（4.8Gbps）带宽提高了80%。这种定制化产品专门针对数据中心和高性能计算场景，满足了对高带宽和低延迟的需求。
移动设备与XR应用	三星电子正在开发LLW DRAM（低延迟宽I/O DRAM），这种新型DRAM通过增加输入/输出端子数量来提升带宽，具有128GB/s的高性能和低延迟特性。LLW DRAM有望应用于端侧AI和下一代XR设备，如苹果的Vision Pro，取代现有的LPDDR产品。这种定制化DRAM能够满足移动设备和XR设备对高性能、低延迟内存的需求。 华邦电子开发的创新型CUBE（Customized Ultra Bandwidth Element，定制化超高带宽元件）技术，旨在大幅提升内存接口带宽，以满足边缘计算平台上快速增长的AI应用需求。CUBE作为一款高带宽、低功耗、紧凑尺寸、极具成本效益，以及可定制化的内存解决方案，可以满足AI应用日益增长的需求，并且可供模组制造商和SoC厂商直接部署。

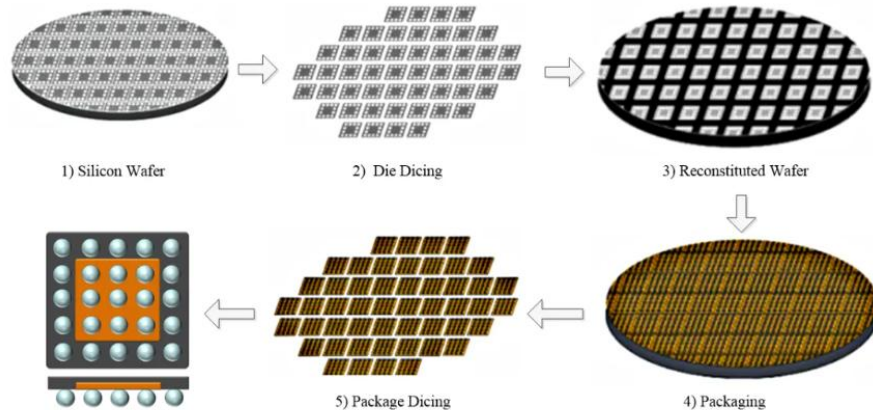
五

先进封装

- 先进封装技术是半导体行业近年来的重要发展方向，旨在通过创新的封装架构和工艺，提升芯片性能、降低功耗、减小尺寸，并优化成本。可以列出的先进封装相关的名称至少有几十种，为了便于区分，我们将先进封装分为两大类：① 基于XY平面延伸的先进封装技术，主要通过RDL进行信号的延伸和互连；② 基于Z轴延伸的先进封装技术，主要是通过TSV进行信号延伸和互连。
- 这里的XY平面指的是Wafer或者芯片的XY平面，这类封装的鲜明特点就是没有TSV硅通孔，其信号延伸的手段或技术主要通过RDL层来实现，通常没有基板，其RDL布线时是依附在芯片的硅体上，或者在附加的Molding上。因为最终的封装产品没有基板，所以此类封装都比较薄，在智能手机中得到广泛的应用。

图表24：基于XY平面延伸的先进封装技术

封装技术	介绍
FOWLP	➢ FOWLP (Fan-out Wafer Level Package)是WLP (Wafer Level Package)晶圆级封装的一种。在WLP技术出现之前，传统封装工艺步骤主要在裸片切割分片后进行，先对晶圆 (Wafer) 进行切割分片 (Dicing)，然后再封装 (Packaging)成各种形式。 ➢ WLP于2000年左右问世，有两种类型：Fan-in (扇入式) 和Fan-Out (扇出式)，WLP晶圆级封装和传统封装不同，在封装过程中大部分工艺过程都是对晶圆进行操作，即在晶圆上进行整体封装 (Packaging)，封装完成后再进行切割分片。因为封装完成后再进行切割分片，因此，封装后的芯片尺寸和裸芯片几乎一致，因此也被称为CSP (Chip Scale Package) 或者WLCSP (Wafer Level Chip Scale Packaging)，此类封装符合消费类电子产品轻、小、短、薄化的市场趋势，寄生电容、电感都比较小，并具有低成本、散热佳等优点。开始WLP多采用Fan-in型态，可称之为Fan-in WLP 或者FIWLP，主要应用于面积较小、引脚数量少的芯片。 ➢ 随着IC工艺的提升，芯片面积缩小，芯片面积内无法容纳足够的引脚数量，因此衍生出Fan-Out WLP 封装形态，也称为FOWLP，实现在芯片面积范围外充分利用RDL做连接，以获取更多的引脚数。 ➢ FOWLP，由于要将RDL和Bump引出到裸芯片的外围，因此需要先进行裸芯片晶圆的划片分割，然后将独立的裸芯片重新配置到晶圆工艺中，并以此为基础，通过批量处理、金属化布线互连，形成最终封装。

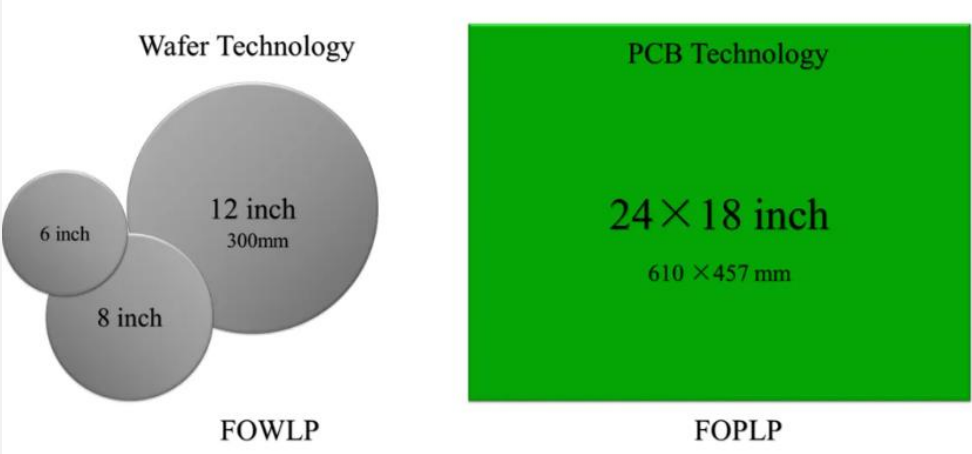


FOWLP封装流程

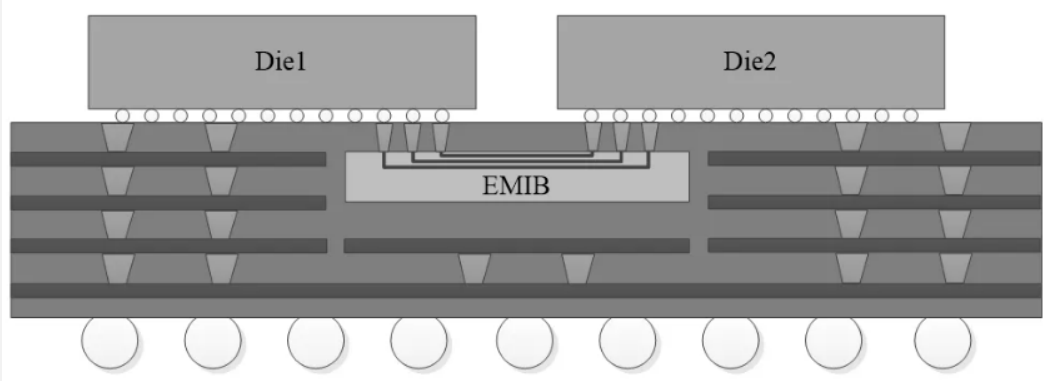
图表24：基于XY平面延伸的先进封装技术（接上表）

封装技术	介绍																						
FOWLP	➢ FOWLP受到很多公司的支持，不同公司有不同的命名方法。 ➢ 无论是采用Fan-in还是Fan-out，WLP晶圆级封装和PCB的连接都是采用倒装芯片形式，芯片有源面朝下对着印刷电路板，可以实现最短的电路路径，这也保证了更高的速度和更少的寄生效应。另一方面，由于采用批量封装，整个晶圆能够实现一次全部封装，成本的降低也是晶圆级封装的另一个推动力量。																						
	<table border="1"> <thead> <tr> <th>Company</th><th>Package</th></tr> </thead> <tbody> <tr> <td>ASE</td><td>eWLB</td></tr> <tr> <td>Amkor Technology(including NANIM)</td><td>eWLB/SWIFT</td></tr> <tr> <td>Deca Technologies</td><td>M-Series</td></tr> <tr> <td>Huatian Technology</td><td>eSiFO</td></tr> <tr> <td>Infineon</td><td>eWLB</td></tr> <tr> <td>JCAP</td><td>ECP</td></tr> <tr> <td>Nepes(developed originally by Freescale)</td><td>RCP</td></tr> <tr> <td>SPIL</td><td>TPI-FO</td></tr> <tr> <td>STATS ChipPAC</td><td>eWLB</td></tr> <tr> <td>TSMC</td><td>InFO-WLP</td></tr> </tbody> </table> 提出扇出晶圆级封装的公司	Company	Package	ASE	eWLB	Amkor Technology(including NANIM)	eWLB/SWIFT	Deca Technologies	M-Series	Huatian Technology	eSiFO	Infineon	eWLB	JCAP	ECP	Nepes(developed originally by Freescale)	RCP	SPIL	TPI-FO	STATS ChipPAC	eWLB	TSMC	InFO-WLP
Company	Package																						
ASE	eWLB																						
Amkor Technology(including NANIM)	eWLB/SWIFT																						
Deca Technologies	M-Series																						
Huatian Technology	eSiFO																						
Infineon	eWLB																						
JCAP	ECP																						
Nepes(developed originally by Freescale)	RCP																						
SPIL	TPI-FO																						
STATS ChipPAC	eWLB																						
TSMC	InFO-WLP																						
InFO	➢ InFO (Integrated Fan-out) 是台积电 (TSMC) 于2017年开发出来的FOWLP先进封装技术，是在FOWLP工艺上的集成，可以理解为多个芯片Fan-Out工艺的集成，而FOWLP则偏重于Fan-Out封装工艺本身。InFO给予了多个芯片集成的空间，可应用于射频和无线芯片的封装，处理器和基带芯片封装，图形处理器和网络芯片的封装。 a) FIWLP b) FOWLP c) InFO FIWLP, FOWLP和InFO对比示意图 ➢ 苹果iPhone处理器早年一直是三星来生产，但台积电却从苹果A11 开始，接连独拿两代iPhone处理器订单，关键之一，就在于台积电全新封装技术InFO，能让芯片与芯片之间直接互连，减少厚度，腾出宝贵的空间给电池或其他零件使用。苹果和台积电的加入改变了FOWLP技术的应用状况，将使市场开始逐渐接受并普遍应用FOWLP (InFO) 封装技术。																						

图表24：基于XY平面延伸的先进封装技术（接上表）

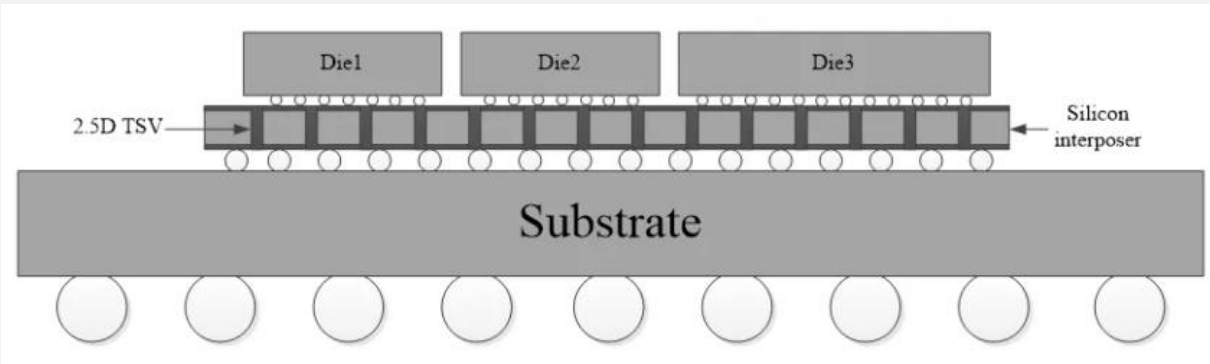
封装技术	介绍
FOPLP	➤ FOPLP（Fan-out Panel Level Package）面板级封装，借鉴了FOWLP的思路和技术，但采用了更大的面板，因此可以量产出数倍于 300 毫米硅晶圆芯片的封装产品。FOPLP技术是FOWLP 技术的延伸，在更大面积的方形载板上进行Fan-Out制程，因此被称为 FOPLP 封装技术，其Panel载板可以采用PCB载板，或者液晶面板用的玻璃载板。目前而言，FOPLP采用了如 24×18英寸（610×457mm）的PCB载板，其面积大约是 300 mm硅晶圆的4 倍，因而可以简单的视为在一次的制程下，就可以量产出 4 倍于300 mm硅晶圆的先进封装产品。和FOWLP工艺相同，FOPLP 技术可以将封装前后段制程整合进行，可以将其视为一次的封装制程，因此可大幅降低生产与材料等各项成本。  Wafer Technology PCB Technology 6 inch 8 inch 12 inch 300mm 24×18 inch 610×457 mm FOWLP FOPLP FOPLP和FOWLP比较 ➤ FOPLP采用了PCB上的生产技术进行RDL的生产，其线宽、线间距目前均大于10um，采用SMT设备进行芯片和无源器件的贴装，由于其面板面积远大于晶圆面积，因而可以一次封装更多的产品。相对FOWLP，FOPLP具有更大的成本优势。目前，全球各大封装业者包括三星电子、日月光均积极投入到FOPLP 制程技术中。

图表24：基于XY平面延伸的先进封装技术（接上表）

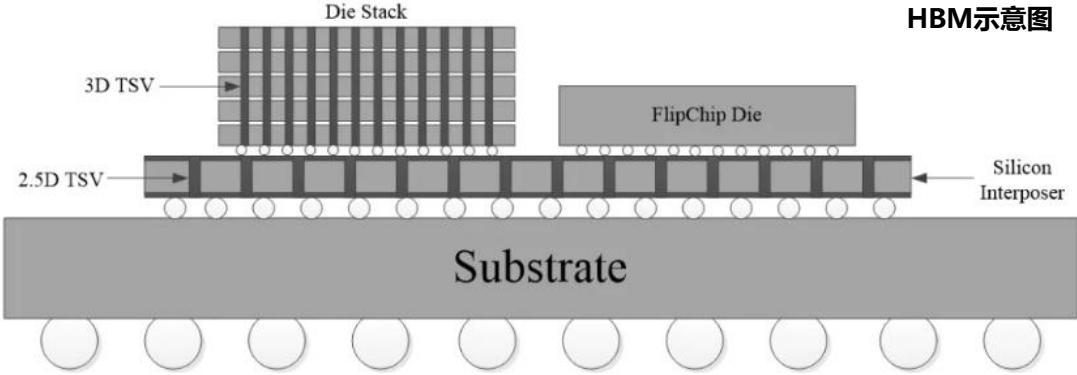
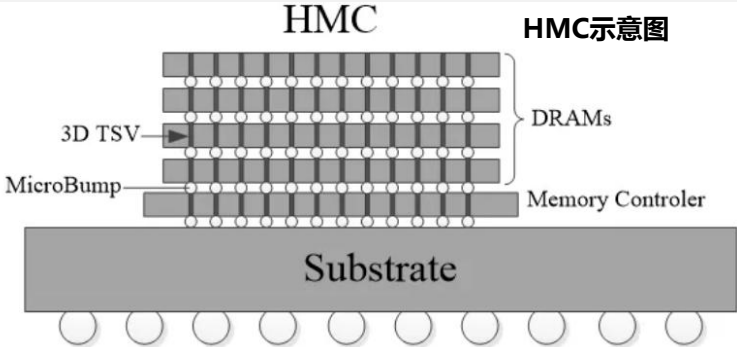
封装技术	介绍
EMIB	➢ EMIB（Embedded Multi-Die Interconnect Bridge）嵌入式多芯片互连桥先进封装技术是由英特尔提出并积极应用的，和前面描述的3种先进封装不同，EMIB是属于有基板类封装，因为EMIB也没有TSV，因此也被划分到基于XY平面延伸的先进封装技术。 ➢ EMIB理念跟基于硅中介层的2.5D封装类似，是通过硅片进行局部高密度互连。与传统2.5封装的相比，因为没有TSV，因此EMIB技术具有正常的封装良率、无需额外工艺和设计简单等优点。传统的SoC芯片，CPU、GPU、内存控制器及IO控制器都只能使用一种工艺制造。采用EMIB技术，CPU、GPU对工艺要求高，可以使用10nm工艺，IO单元、通讯单元可以使用14nm工艺，内存部分则可以使用22nm工艺，采用EMIB先进封装技术可以把三种不同工艺整合到一起成为一个处理器。  EMIB示意图 ➢ 和硅中介层（interposer）相比，EMIB硅片面积更微小、更灵活、更经济。EMIB封装技术可以根据需要将CPU、IO、GPU甚至FPGA、AI等芯片封装到一起，能够把10nm、14nm、22nm等多种不同工艺的芯片封装在一起做成单一芯片，适应灵活的业务的需求。通过EMIB方式，KBL-G平台将英特尔酷睿处理器与AMD Radeon RX Vega M GPU整合在一起，同时具备了英特尔处理器强大的计算能力与AMD GPU出色的图形能力，并且还有着极佳的散热体验。

- 基于Z轴延伸的先进封装技术主要是通过TSV进行信号延伸和互连，TSV可分为2.5D TSV和3D TSV，通过TSV技术，可以将多个芯片进行垂直堆叠并互连。
- 在3D TSV技术中，芯片相互靠得很近，所以延迟会更少，此外互连长度的缩短，能减少相关寄生效应，使器件以更高的频率运行，从而转化为性能改进，并更大程度的降低成本。
- TSV技术是三维封装的关键技术，包括半导体集成制造商、集成电路制造代工厂、封装代工厂、新兴技术开发商、大学与研究所以及技术联盟等研究机构都对TSV的工艺进行了多方面的研发。
- 虽然基于Z轴延伸的先进封装技术主要是通过TSV进行信号延伸和互连，但RDL同样也是不可或缺的，例如，如果上下层芯片的TSV无法对齐时，就需要通过RDL进行局部互连。

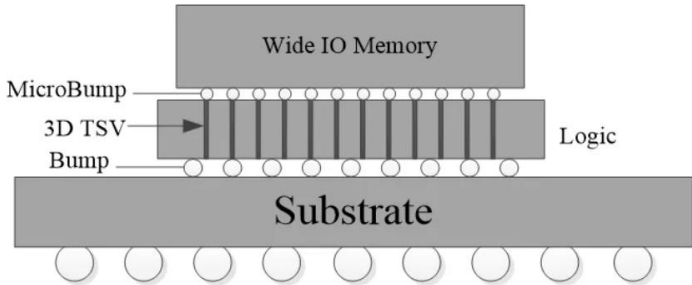
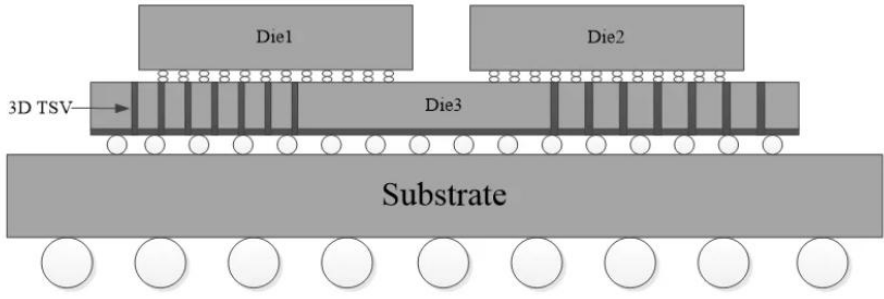
图表25： 基于Z轴延伸的先进封装技术

封装技术	介绍
CoWoS	➢ CoWoS (Chip-on-Wafer-on-Substrate) 是台积电推出的 2.5D封装技术，CoWoS是把芯片封装到硅转接板（中介层）上，并使用硅转接板上的高密度布线进行互连，然后再安装在封装基板上，如下图所示。  CoWoS示意图 ➢ CoWoS和前面讲到的InFO都来自台积电，CoWoS有硅转接板Silicon Interposer，InFO则没有。CoWoS针对高端市场，连线数量和封装尺寸都比较大。InFO针对性价比市场，封装尺寸较小，连线数量也比较少。台积电2012年就开始量产CoWoS，通过该技术把多颗芯片封装到一起，通过Silicon Interposer高密度互连，达到了封装体积小，性能高、功耗低，引脚少的效果。

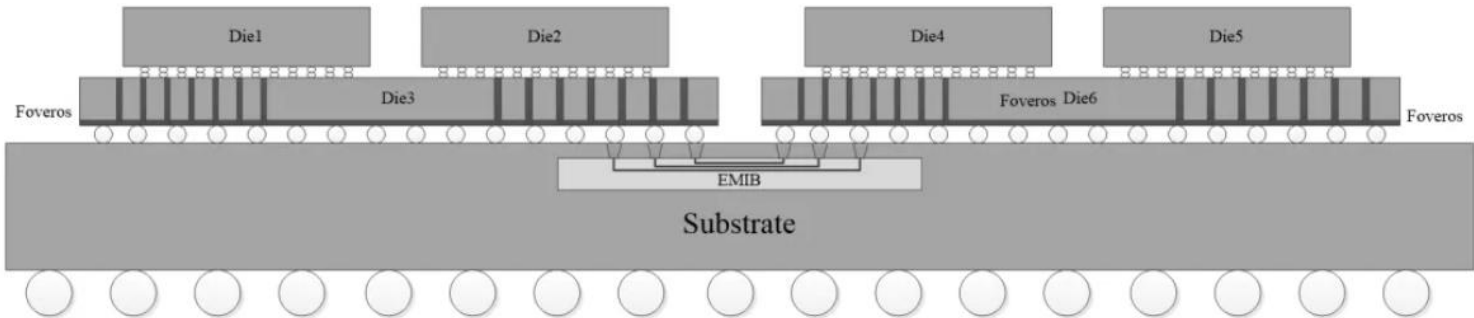
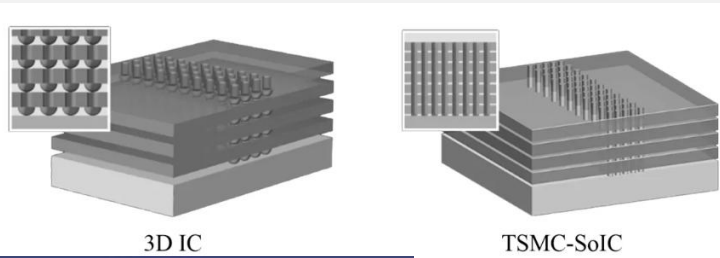
图表25： 基于Z轴延伸的先进封装技术（接上表）

封装技术	介绍
HBM	➢ HBM (High-Bandwidth Memory) 高带宽内存，主要针对高端显卡市场。HBM使用了3D TSV和2.5D TSV技术，通过3D TSV把多块内存芯片堆叠在一起，并使用2.5D TSV技术把堆叠内存芯片和GPU在载板上实现互连。  HBM示意图
HMC	➢ HMC (Hybrid Memory Cube) 混合存储立方体，其标准由美光主推，目标市场是高端服务器市场，尤其是针对多处理器架构。HMC使用堆叠的DRAM芯片实现更大的内存带宽。另外HMC通过3D TSV集成技术把内存控制器 (Memory Controller) 集成到DRAM堆叠封装里。  HMC示意图 ➢ 对比HBM和HMC，两者都是将DRAM芯片堆叠并通过3D TSV互连，并且其下方都有逻辑控制芯片，两者的不同在于：HBM通过Interposer和GPU互连，而HMC则是直接安装在Substrate上，中间缺少了Interposer和2.5D TSV。在HMC堆叠中，3D TSV的直径约为5~6um，数量超过了2000+，DRAM芯片通常减薄到50um，之间通过20um的MicroBump将芯片相连。以往内存控制器都做在处理器里，所以在高端服务器里，当需要使用大量内存模块时，内存控制器的设计非常复杂。现在把内存控制器集成到内存模块内，则内存控制器的设计大大简化。此外，HMC使用高速串行接口 (SerDes) 来实现高速接口，适合处理器和内存距离较远的情况。

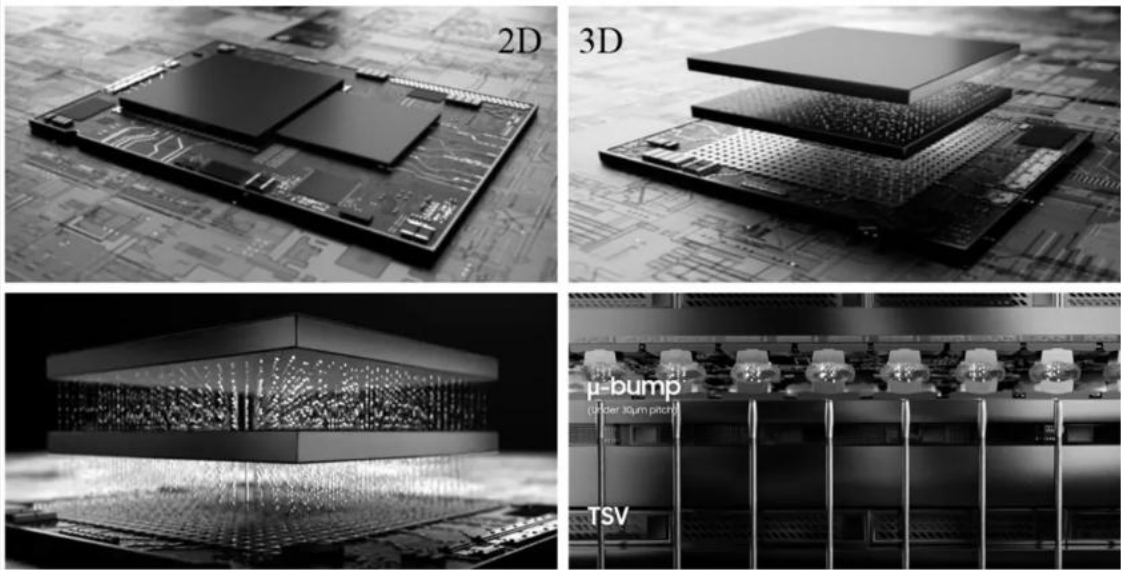
图表25： 基于Z轴延伸的先进封装技术（接上表）

封装技术	介绍
Wide-IO	➤ Wide-IO (Wide Input Output) 宽带输入输出技术由三星主推，目前已经到了第二代，可以实现最多512bit的内存接口位宽，内存接口操作频率最高可达1GHz，总的内存带宽可达68GBps，是DDR4接口带宽（34GBps）的两倍。Wide-IO通过将Memory芯片堆叠在Logic芯片上来实现，Memory芯片通过3D TSV和Logic芯片及基板相连接，如下图所示。  Wide-IO示意图 ➤ Wide-IO具备TSV架构的垂直堆叠封装优势，有助打造兼具速度、容量与功率特性的移动存储器，满足智能手机、平板电脑、掌上型游戏机等行动装置的需求，其主要目标市场是要求低功耗的移动设备。
Foveros	➤ 除了EMIB之外，Intel还推出了Foveros有源板载技术。在Intel的技术介绍中，Foveros被称作3D Face to Face Chip Stack for heterogeneous integration，三维面对面异构集成芯片堆叠。EMIB与Foveros的区别在于前者是2D封装技术，而后者则是3D堆叠封装技术，与2D的EMIB封装方式相比，Foveros更适用于小尺寸产品或对内存带宽要求更高的产品。EMIB和Foveros在芯片性能、功能方面的差异不大，都是将不同规格、不同功能的芯片集成在一起，但在体积、功耗等方面，Foveros 3D堆叠的优势就显现了出来。Foveros每比特传输的数据的功率非常低，Foveros技术要处理的是Bump间距减小、密度增大以及芯片堆叠技术。  Foveros示意图 ➤ 从Meteor Lake开始，Intel将Foveros技术引入客户端产品，如第13代酷睿处理器。这些处理器将多种功能(如CPU、GPU、PCH)整合到一块芯片上，Foveros封装产能提升四倍。

图表25： 基于Z轴延伸的先进封装技术（接上表）

封装技术	介绍
Co-EMIB (Foveros + EMIB)	➢ Co-EMIB (Combined Embedded Multi-die Interconnect Bridge) 是英特尔推出的一种先进封装技术，结合了2.5D的EMIB和3D的Foveros封装技术，EMIB主要是负责横向的连结，让不同内核的芯片像拼图一样拼接起来，Foveros则是纵向堆栈，就像盖高楼一样，每层楼都可以有完全不同的设计。该技术可以将多个3D Foveros芯片通过EMIB拼接在一起，以制造更大的芯片系统，实现高性能、低功耗和高带宽。  Wide-IO示意图 ➢ 达成Co-EMIB技术的关键，是ODI (Omni-Directional Interconnect) 全向互连技术。ODI具有两种不同型态，除了打通不同层的电梯型态连接外，也有连通不同立体结构的天桥，以及层之间的夹层，让不同的芯片组合可以有极高的弹性。
SoIC	➢ 台积电的SoIC (System on Integrated Chips) 集成片上系统封装技术正在快速发展，其3D堆叠技术的凸块间距预计到2027年将从目前的9μm缩小至3μm。 ➢ SoIC是一种创新的多芯片堆栈技术，能对10纳米以下的制程进行晶圆级的集成。SoIC-X的特点是没有凸点 (no-Bump) 的键合结构，因此具有有更高的集成密度和更佳的运行性能。 ➢ SoIC包含CoW (Chip-on-wafer) 和WoW (Wafer-on-wafer) 两种技术形态，从TSMC的描述来看，SoIC就一种WoW晶圆对晶圆或CoW芯片对晶圆的直接键合 (Bonding) 技术，属于Front-End 3D技术 (FE 3D)，而前面提到的InFO和CoWoS则属于Back-End 3D技术 (BE 3D)。 ➢ SoIC和3D IC的制程有些类似，SoIC的关键在于实现没有凸点的接合结构，并且其TSV的密度也比传统的3D IC密度更高，直接通过极微小的TSV来实现多层芯片之间的互联。如右图所示是3D IC和SoIC两者中TSV密度和Bump尺寸的比较。可以看出，SoIC的TSV密度要远远高于3D IC，同时其芯片间的互联也采用no-Bump的直接键合技术，芯片间距更小，集成密度更高。  3D IC TSMC-SoIC

图表25： 基于Z轴延伸的先进封装技术（接上表）

封装技术	介绍
X-Cube	➤ X-Cube (eXtended-Cube) 是三星宣布推出的一项3D集成技术，可以在较小的空间中容纳更多的内存，并缩短单元之间的信号距离。X-Cube用于需要高性能和带宽的工艺，例如5G，人工智能以及可穿戴或移动设备以及需要高计算能力的应用中。X-Cube利用TSV技术将SRAM堆叠在逻辑单元顶部，可以在更小的空间中容纳更多的存储器。从X-Cube技术展示图可以看到，不同于以往多个芯片2D平行封装，X-Cube 3D封装允许许多枚芯片堆叠封装，使得成品芯片结构更加紧凑。芯片之间采用了TSV技术连接，降低功耗的同时提高了传输的速率。该技术应用于最前沿的5G、AI、AR、HPC、移动芯片以及VR等领域。  X-Cube示意图 ➤ X-Cube技术大幅缩短了芯片间的信号传输距离，提高数据传输速度，降低功耗，并且还可以按客户需求定制内存带宽及密度。目前X-Cube技术已经可以支持7nm及5nm工艺，三星将继续与全球半导体公司合作，将该技术部署在新一代高性能芯片中。

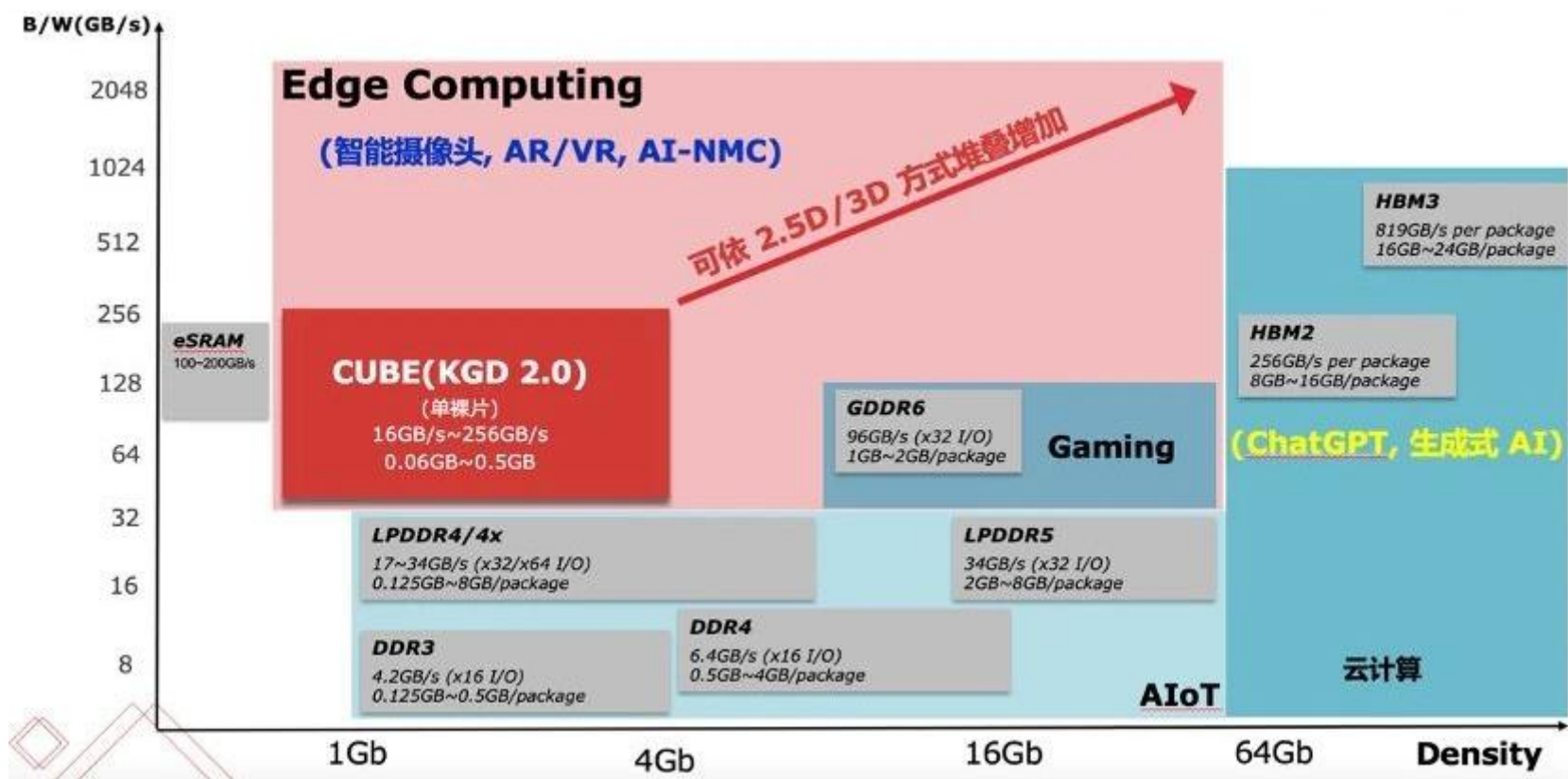


定制化存储：华邦CUBE介绍

CUBE: 用于边缘计算且具备可扩展性

- 华邦电子开发的创新型CUBE (Customized Ultra Bandwidth Element, 定制化超高带宽元件) 技术, 旨在大幅提升内存接口带宽, 以满足边缘计算平台上快速增长的AI应用需求。CUBE作为一款高带宽、低功耗、紧凑尺寸、极具成本效益, 以及可定制化的内存解决方案, 可以满足AI应用日益增长的需求, 并且可供模组制造商和SoC厂商直接部署。

图表26: CUBE用于边缘计算且具备可扩展性



- CUBE是客制化的高宽带存储芯片3D TSV DRAM, 专门为边缘AI运算装置所设计的存储架构, 利用3D堆叠技术并结合异质键合技术以提供高带宽、低功耗、单颗256Mb至8Gb的存储芯片。
- **架构:** CUBE是将SoC裸片置上, DRAM裸片置下, 可以省去SoC中的TSV工艺, 进而降低了SoC裸片的尺寸与成本。同时, 3D DRAM TSV工艺可以将SoC信号引至外部, 使它们成为同一颗芯片, 进一步缩减了封装尺寸。而因为SoC裸片在上方也可以有比较好的散热效果。
- **制造:** 据了解, 这个专案当时由联电推动, 目标是锁定边缘运算AI应用在穿戴式装置、家用和工业物联网、安全和智慧基础设施等, 提供中高阶算力、可客制的存储模组和较低功耗需求的解决方案。联电负责CMOS晶圆制造和晶圆对晶圆混合封装技术, 华邦电导入客制化CUBE架构, 智原提供全面的3D先进封装一站式服务, 以及存储IP和ASIC小芯片设计服务, 日月光则提供晶圆切割、封装和测试服务, 另外还有Cadence负责晶圆对晶圆设计流程, 提取TSV特性和签核认证。

图表27: HBM和CUBE对比

	HBM1	HBM2	HBM2E	HBM3	HBM3E	HBM4	CUBEX
堆叠层数	4层	8层	8层/12层	8层/12层	8层/12层	12层/16层	4层
I/O传输速率	1Gb/s	2Gb/s	3.6Gb/s	6.4Gb/s	9.2Gb/s	9.2Gb/s	2Gb/s
整体封装带宽	128GB/s	256GB/s	460GB/s	819GB/s	1.2TB/s	2.4TB/s	1TB/s
封装容量	4GB	8GB	16GB/24GB	16GB/24GB	24GB/36GB	36/48GB	2-4GB
数据位宽	1024	1024	1024	1024	1024	2048	4096
功耗	6pJ/bit	<5pJ/bit	<5pJ/bit	<4pJ/bit	<4pJ/bit	<4pJ/bit	<1pJ/bit
应用	数据中心中的AI/机器学习, 高性能计算, 加速器系统						SoC L4缓存, AI-ISP, 可穿戴设备

资料来源: Winbond, 问芯, 中邮证券研究所

请参阅附注免责声明

更多免费行业报告

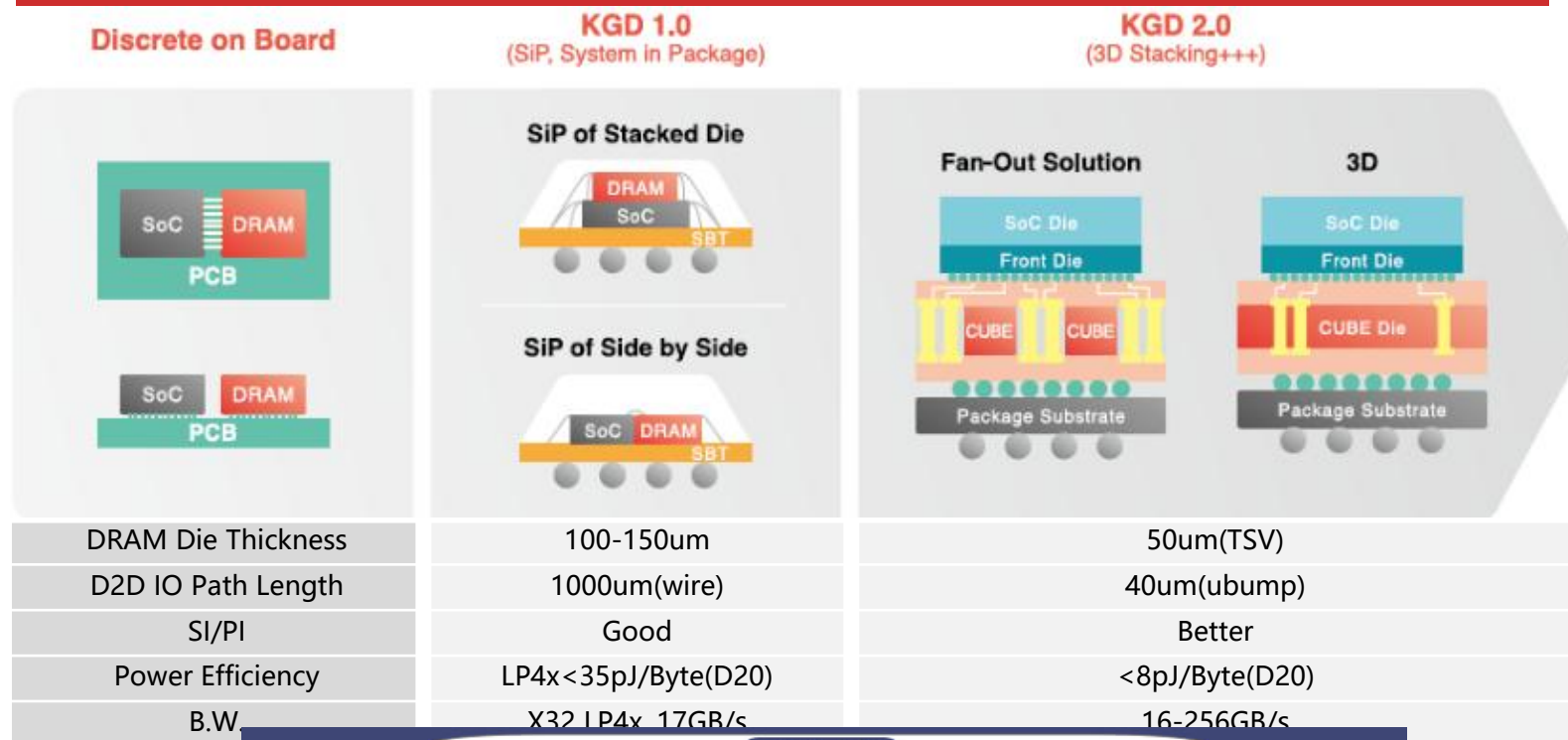
并购家

www.ipoiipo.cn

CUBE: 演进至KGD 2.0 (3D Stacking)

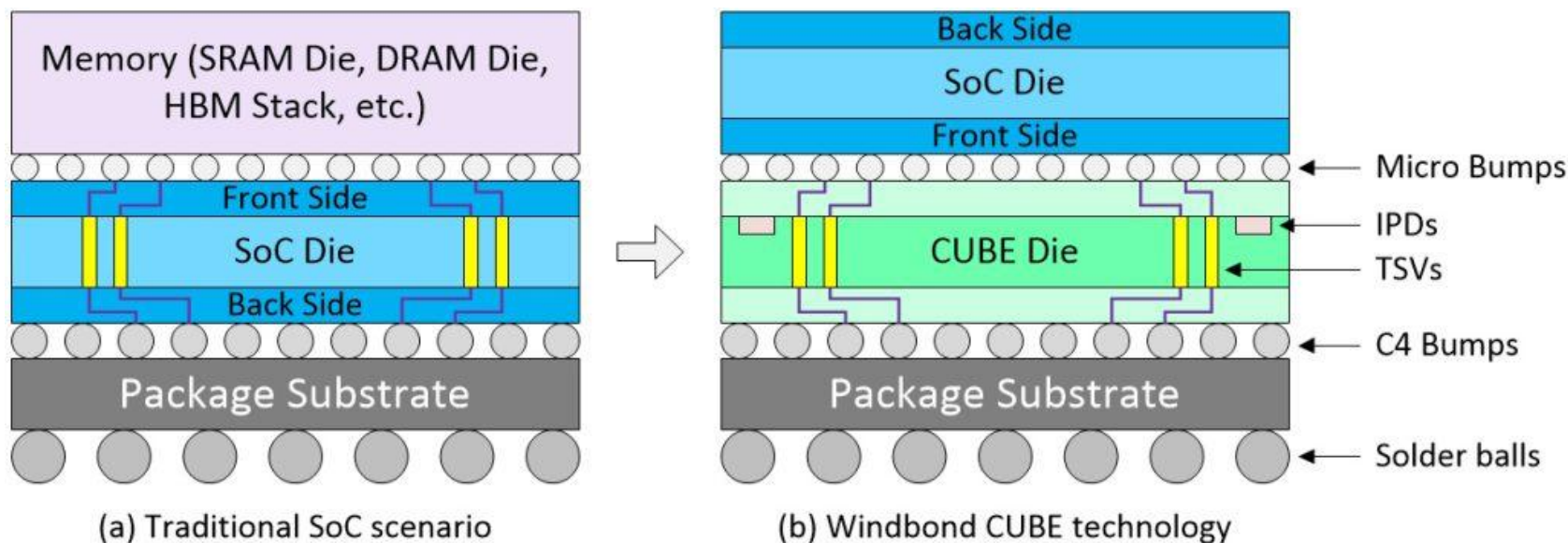
- **演进:** CUBE 的超高带宽实现和 HBM 比较相似, 主要是通过大量并行的 IO 口, 以相对较低的频率达成。大量 IO 连接也决定了 CUBE 需以 3D 堆叠的方式, 藉由 TSV (硅过孔) 与 SoC die 互联。所以华邦与主芯片厂之间的合作, 也从过去的 KGD 1.0 (Known Good Die) 升级到了 KGD 2.0。
- **KGD 1.0:** 指芯片SiP合封, 即DRAM和SOC芯片堆叠或并排封装, 这种模式在业界已成熟多年, 尤其在中小容量DRAM产品 (如1Gbit以下) 的封装中, 具有低成本和小占板面积的优势。
- **KGD 2.0:** 在KGD 2.0模式下, 华邦销售测试好的wafer (晶圆) 给客户, 由客户选择封装厂进行封装。华邦电子产品总监朱迪指出, 未来商业模式中封装厂的角色将更重要。华邦将继续专注于提供DRAM wafer, 并可能提供更多服务, 如测试等。

图表28: 华邦电子KGD 2.0演进历程



- **传统SOC封装：** SoC设计人员以尽可能小的封装为目标，传统解决方案(a)是将memory翻转过来，并将其装在SoC的正面。反过来，这要求SoC采用TSV（硅通孔），这会消耗宝贵的硅空间。
- **华邦电子CUBE：** CUBE(b)技术中，SoC被翻转并连接到CUBE晶片上，1) SoC芯片置于顶部可提供更好的功耗；2) SoC 芯片中没有TSV可以降低成本、复杂性和面积；3) CUBE可以包括硅电容器(Si-Cap)等集成无源器件(IPD)，有助于为SoC提供稳定的电源，而无需无数的分立多层陶瓷电容器(MLCC)，附加值是CUBE充当DRAM+硅电容+硅中介的角色存在——故称其为“element”。每个内核SRAM单元需要6个晶体管，每个核心DRAM单元只需要1个晶体管。因此，使用相对便宜的20nm工艺，CUBE的密度是采用相对昂贵的14nm工艺实现的SRAM芯片的5倍。

图表29：传统SOC封装与华邦电子CUBE技术的比较



- CUBE通过增加I/O数量、提高数据速度、支持TSV（可选）、提供散热优秀的3D架构，解决了传统内存IC和模组解决方案的痛点。

图表30：当前内存市场存在的问题以及CUBE的主要特性

当前内存市场存在的问题	CUBE的主要特性	
为满足运行AI应用的系统需求，内存芯片和模组需要满足多方面的性能需求，不仅仅是高带宽，包括低能耗、合理的尺寸，以及散热管理。	较小尺寸	基于D20标准的1-8Gb/die 产品，以及灵活的设计和3D堆叠选择，使得CUBE能够适应更小的外形尺寸。TSV的引入也进一步提高了性能，改善了信号完整性、电源完整性和散热性能。
现有内存解决方案在带宽方面存在限制，直接影响了系统的性能。IC引脚数量、数据传输速率和内存总线宽度等物理特性，在决定接口带宽方面也起着重要作用。	高经济效益、高带宽	基于D20工艺的CUBE可以设计为1-8Gb/die 容量，基于D16工艺的为16Gb/die 容量。非TSV和TSV堆叠均可用，这为各种应用提供了优化内存带宽的灵活性。
内存芯片和模块设计师还须应对设计和制造工艺的限制，这些限制会影响所提供的带宽。此外，随着速度提升，信号完整性也是一个关键问题，因为衰减、干扰和反射等因素也会限制可实现的带宽。	SoC 芯片尺寸减小	TSV技术以及uBump/ 混合键合可降低功耗并节省SoC设计面积，从而实现高效且极具成本效益的解决方案。利用TSV实现高效的3D堆叠，简化了与先进封装技术的集成难度。通过减小芯片尺寸，CUBE能以更短电源路径以及更紧凑、更轻巧的设计来降低器件成本、提高能效。
另一方面，增加带宽可能会造成功耗上升或效率降低，这会给散热管理带来很大挑战，并影响依靠电池供电的边缘设备运行。此外，一些解决方案会导致内存模块体积增大，限制其在紧凑型设备中的应用。	节省电耗	CUBE具有出色的能效，在D20工艺中功耗低于1pJ/bit，能够延长设备运行时间、优化电源消耗。
随着用户日益依赖人工智能应用，特别是依赖大模型的应用，带宽、功耗和外形尺寸等限制将对所有内存技术构成更大挑战。这些不断涌现并快速发展的工作负载需要更高效节能的计算能力来满足需求。	性能卓越	CUBE的IO速度于1K I/O可高达2Gbps，提供从16GB/s 至256GB/s 的总带宽。通过这种方式，CUBE能够确保带来高于行业标准的性能提升，并通过uBump或混合键合增强电源和信号完整性。

- 华邦的CUBE解决方案主要面向低功耗、高带宽，以及中低容量的内存需求，适合于边缘计算和生成式AI等应用。
- 例如在AI-ISP架构中，如下图所示，灰色部分属于神经网络处理器（NPU），如果AI-ISP要实现大算力，需要很大的带宽，或者是SPRAM。但是在AI-ISP上使用SPRAM的成本非常高，不太可行。如果使用LPDDR4就需要4-8颗，无论是合封还是外置，成本同样相当高昂。此外，还有可能会用到传输速度为4266Mhz的高速LPDDR4，而这样的产品需要依赖7nm或12nm的先进制程工艺生产。
- 华邦的CUBE解决方案可以允许客户使用成熟制程（例如28、22nm）的SoC，获得类似的高速带宽。华邦的CUBE解决方案可以通过多个I/O（256或者512个）结合28nm SoC提供的500MHz的运行频率，以此实现更高带宽，带宽最高可增至256GB/s。不仅如此，华邦在未来可能会和客户探讨64GB/s带宽的合作，I/O数可以减少，裸片的尺寸也会进一步缩小。

图表31：华邦的CUBE在AI-ISP架构的替代性解决方案

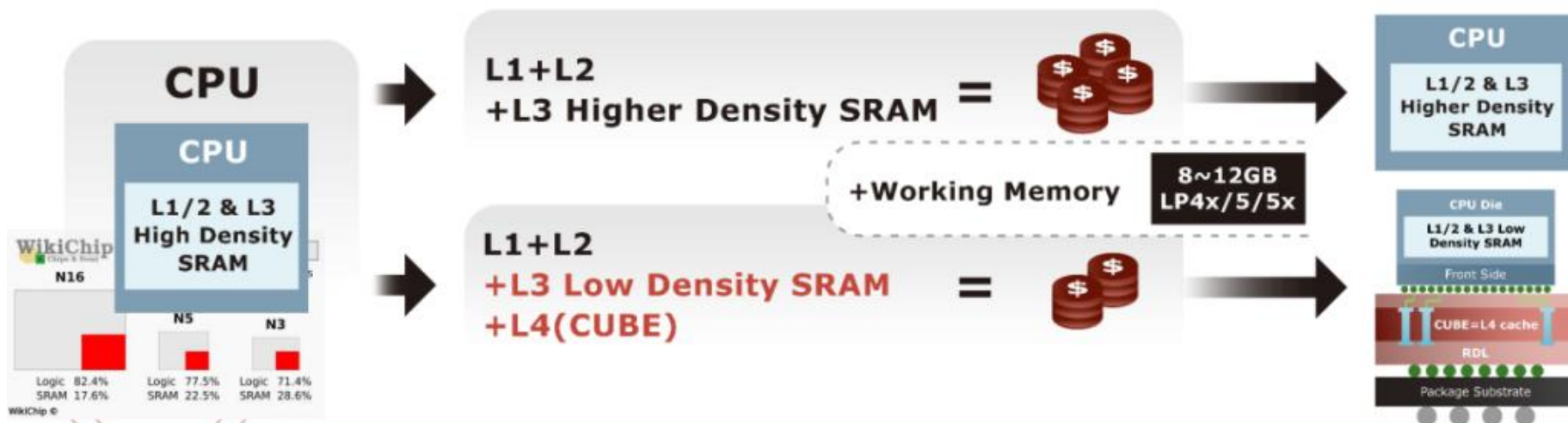


资料来源：电子发烧友，中邮证券研究所

请参阅附注免责声明

- 随着CPU高速运算需求对制程的要求越来越高，可以看到16nm、7nm、5nm到3nm的CPU，SRAM占比（如下图中红色部分所示）并不会同比例缩小，因此当需要实现AI运算或者进行高速运算的情况下，就需要把L3的缓存SRAM容量加大，即便可以使用堆叠的方式达到几百MB，也会导致高昂的成本。
- 华邦的方案是把L3缓存缩小，转而使用L4缓存的CUBE解决方案。L4缓存之所以被称作L4，首先是因为它的延迟（Latency）会比L3的稍长。华邦电子次世代内存产品营销企划经理曾一峻表示，为了克服这个问题，可以采用多BANK的方式（multibank per channel），来获得更好的存取效率。第二个方式是将重写（rewrite）IO分开，这是一个比较类SRAM的方式，缩短运行时间，即以某些比较特殊的架构进行产品修正，会针对客户的一些特殊需求和应用场景进行定制化调配，缩短L4缓存的延迟。同时，AI模型在某些情况下还是需要外置一定容量的内存，例如在某些边缘计算的场景下会需要8-12GB的LPDDR4或者是LPDDR5，因此也可以外挂高容量的工作内存（Working Memory）。
- 综上所述，CUBE可以允许使用成熟制程，以降低SoC成本、减小芯片功耗以及获得高带宽。

图表32：华邦的CUBE方案可作为L4级缓存用于边缘计算



资料来源：电子发烧友，中邮证券研究所

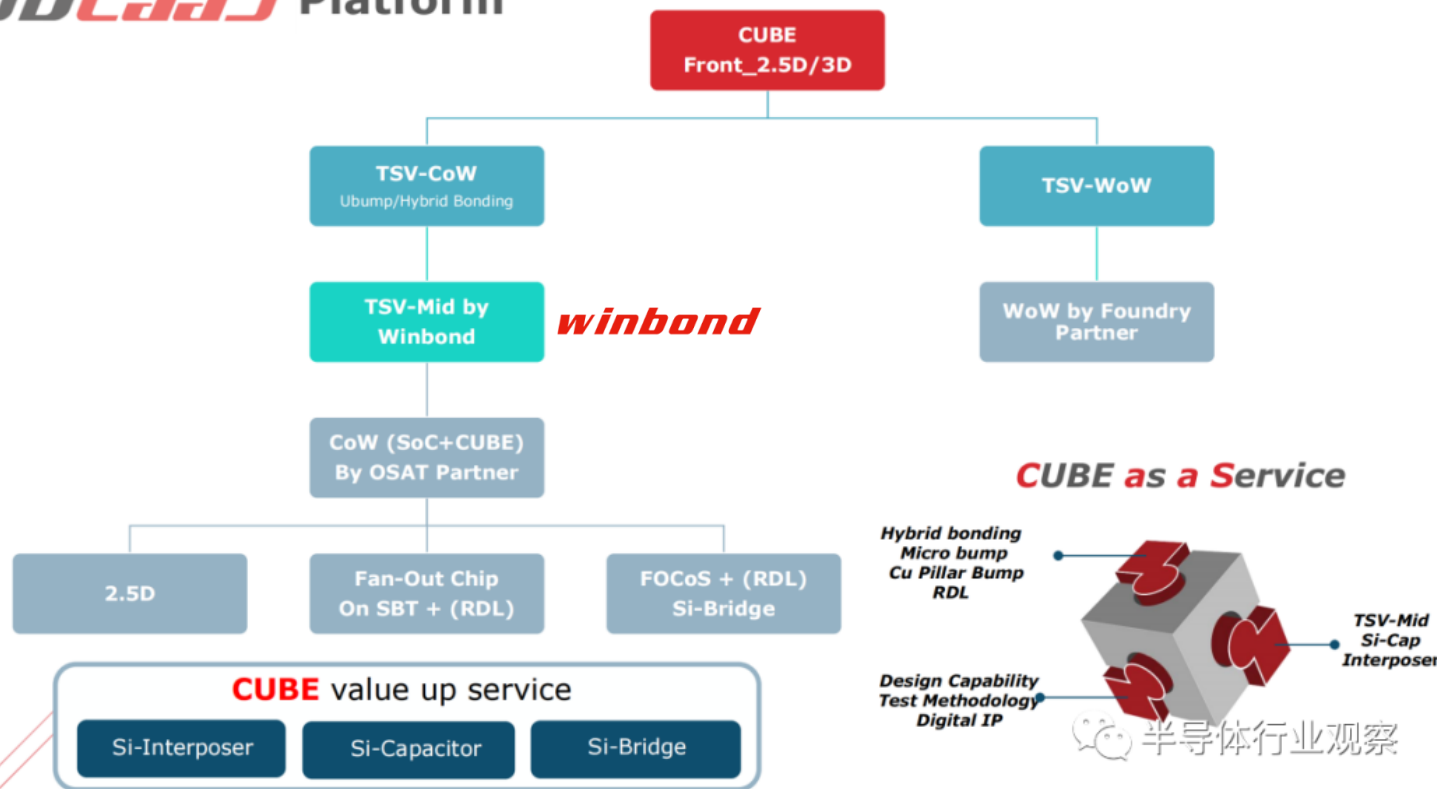
请参阅附注免责声明

3DCaaS (3D CUBE as a Service)一站式服务平台

- 在早前宣布加入UCle联盟的时候华邦表示，公司将提供3DCaaS (3D CUBE as a Service) 一站式服务平台，为客户提供领先的标准化产品解决方案。他们指出，通过此平台，客户不仅可以获得 3D TSV DRAM (又名 CUBE) KGD 内存芯片和针对多芯片设备优化的 2.5D/3D 后段工艺 (采用 CoW/WoW技术)，还可获取由华邦的平台合作伙伴提供的技术咨询服务。这意味着客户可轻松获得完整且全面的 CUBE 产品支持，并享受 Silicon-Cap、interposer 等技术的附加服务。其中，CUBE正是华邦3DCaaS服务的核心之一。

图表33：华邦的3DCaaS一站式服务平台

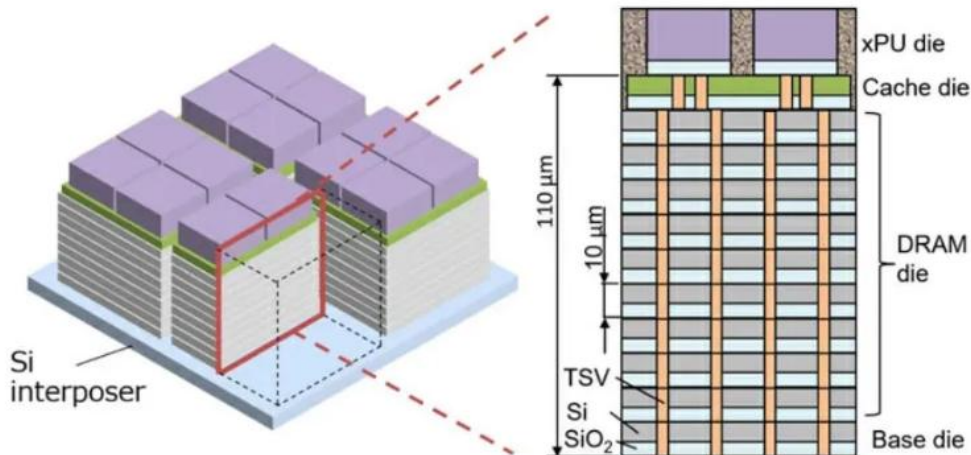
3DCaaS Platform



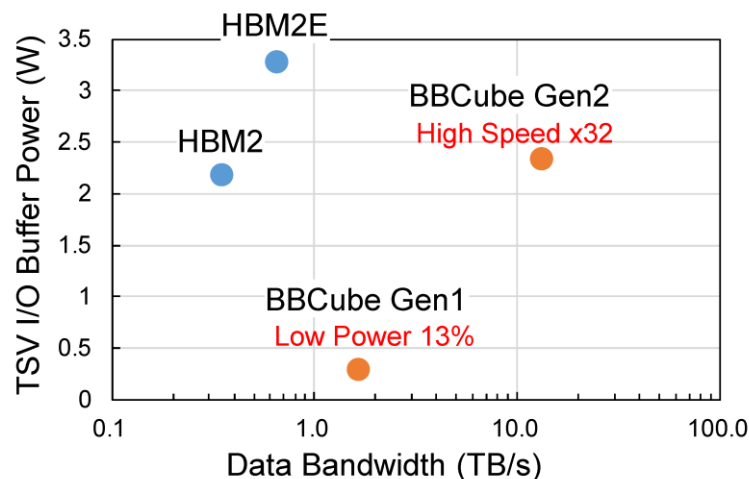
制造：由联电推动，目标是锁定边缘运算AI应用在穿戴式装置、家用和工业物联网、安全和智慧基础设施等，提供中高阶算力、可客制的存储模组和较低功耗需求的解决方案。联电负责CMOS晶圆制造和晶圆对晶圆混合封装技术，华邦导入客制化CUBE架构，智原提供全面的3D先进封装一站式服务，以及存储IP和ASIC小芯片设计服务，日月光则提供晶圆切割、封装和测试服务，另外还有Cadence负责晶圆对晶圆设计流程，提取TSV特性和签核认证。

- 在IEEE 2023年超大规模集成电路技术和电路研讨会的研究中，东京工业大学Takayuki Ohba教授及其同事提出了一项名为Bumpless Build Cube 3D (BBCube 3D)的技术。这项技术将TSV堆叠内存与XPU垂直集成，可提升带宽、减少功耗。
- BBCube 3D可以实现更好的散热，因为TSV可以将DRAM的热量传导到基板上。此外，由于TSV连接较短，无凹凸互连技术可以提高传输速度，并且由于TSV间距更细和无凹凸TSV互连的阻抗更低，因此可以增加每个芯片的TSV密度。

图表34: BBCube 3D 的堆叠架构



图表35: HBM和BBCube的数据带宽和TSV I/O功耗比较



Number of TSV and Transmission speed
 HBM2: 1024 TSV, 2.6 Gb/s/TSV
 HBM2E: 1024 TSV, 5.0 Gb/s/TSV
 BBCube Gen1: 1024 x 16 TSV, 800-Mb/s/TSV
 BBCube Gen2: 1024 x 128 TSV, 800-Mb/s/TSV
 Number of stacking levels is 8
 HBM I/O power is estimation



相关标的

兆易创新：积极开拓定制化存储方案

- **设立青耘科技推动定制化存储方案：**2024年7月，公司披露与关联方共同投资设立控股子公司暨关联交易的公告，公司拟以自有资金出资2,100万元，占控股子公司北京青耘科技有限公司(标的公司)注册资本约77.78%。标的公司设立后，公司计划将其作为针对定制化存储解决方案等创新性业务的子公司进行孵化，在保证公司现有业务稳定发展的基础上，投资创新技术领域，可推动公司积极开拓包括定制化存储方案在内的新技术、新业务、新市场和新产品。根据广东省智能科学（略）高性能存储IP采购项目中标结果公告(2024年12月23日)，青耘获得大幅领先的技术得分，整体排名第一。
- **定制化DRAM和标准DRAM的区别：**公司有比较齐全的利基存储工艺平台，包括NOR Flash、SLC NAND Flash以及DRAM，在市场开拓的过程中有客户提出非标准接口存储器产品的需求。应此需求，公司决定开展定制化方案业务。定制化存储器产品不是标准接口，也不是通用产品，项目开发周期长，业绩释放的周期也较慢，因此设立青耘科技并组建专门的团队来服务客户。与标准接口的存储品产品相比，定制化存储产品的接口、容量等取决于客户对其具体应用的差异化需求。应用场景包括IoT、智能终端等，产品定义来自客户产品的规格要求，通常客户会综合带宽、功耗等指标进行选择和定义。
- **与燧原成立光羽芯辰聚焦端侧大模型芯片：**2024年7月，燧原科技与兆易创新携手成立了端侧大模型芯片公司光羽芯辰，双方各持股15%，该公司聚焦于解决端侧大模型面临的存储带宽和容量问题，通过创新的3D堆叠方案，将燧原的AI技术专长与兆易的DRAM技术深度融合，旨在打造性能卓越、成本更低的定制化AI芯片。

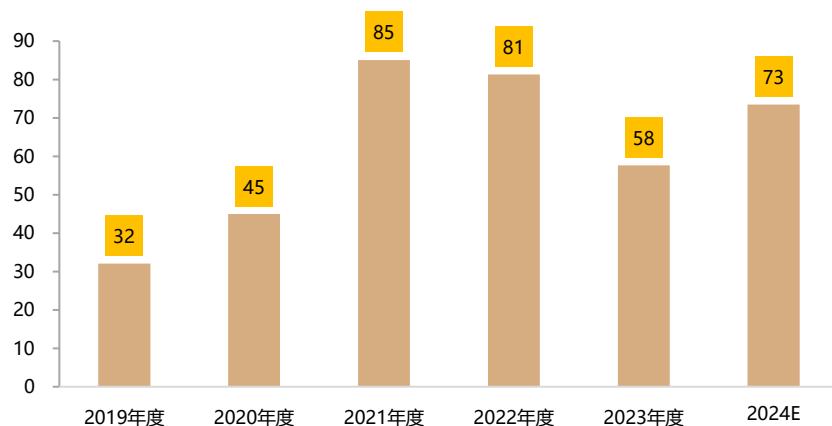
图表36：广东省智能科学与技术研究院高性能存储IP采购项目合同包1(高性能存储IP)

采购标的	服务范围
存储晶圆货物	1. DRAM存储峰值带宽不小于10TB/s； 2. DRAM存储容量不小于20GB；
高性能存储IP设计+存储IP集成服务	高性能存储IP设计： 1. 交付存储控制器IP代码需采用可综合硬件描述语言实现； 2. 交付存储接口IP仿真模型，需兼容主流验证环境； 3. 交付数据表和应用文档，包括性能数据、使用说明、集成说明等； 4. 交付物理设计所需文件（LEF, lib, GDSII等格式文件）； 5. 提供存储IP集成所需的各类交付文件和文档； 存储IP集成服务： 1. 支持存储IP集成与客户的SoC项目中； 2. 及时响应存储IP集成中的各类技术问题； 3. 提供存储IP集成后的DRAM存储方案；
合封好的样片	合封的输入输出触点距不大于5um

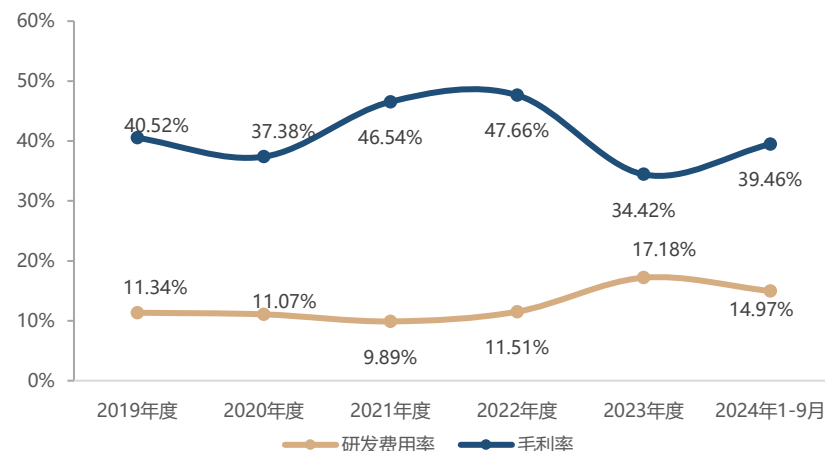
兆易创新：积极开拓定制化存储方案

- **收购苏州赛芯，增强模拟产品能力。**公司拟与石溪资本、合肥国投、合肥产投共同以现金方式收购苏州赛芯70%的股份，且由于前述各联合收购方与公司的表决权委托或一致行动安排，公司将在此次交易完成后成为苏州赛芯的控股股东。此次交易是推动公司模拟战略的重要举措，标的公司的主要产品包括锂电池保护芯片、电源管理芯片等，产品在封装尺寸、产品性能、产品稳定性、产品成本等方面均具有一定竞争力。通过本次收购，公司可进一步增强模拟团队实力，提升电池管理相关技术储备，打开新的成长空间。
- **调整募投项目发展方向及资金投入，前瞻布局LPDDR5和汽车MCU。**1) DRAM：公司根据DRAM产品市场需求变化、产品技术迭代变化，拟将募投项目“DRAM芯片研发及产业化项目”的用途从原有开发四种产品DDR3、DDR4、LPDDR3和LPDDR4调整为DDR3、DDR4、LPDDR和LPDDR5，预计LPDDR5产品将在2029年或2030年进入小容量产品市场，与公司募投项目周期相匹配，助力公司未来发展。2) 汽车MCU：公司本次新增募投项目“汽车电子芯片研发及产业化项目”，公司将完善MCU产品布局，提高高端MCU产品的研发能力，助力车规MCU行业发展，打破国外厂商垄断，进一步扩大市场空间。

图表37：2019-2024E兆易创新收入情况（亿元）



图表38：2019-2024年1-9月兆易创新毛利率、研发费用率



资料来源：iFind，兆易创新公告，中邮证券研究所

请参阅附注免责声明

瑞芯微：持续推进NPU助力端侧AI

- **持续推进NPU助力端侧AI**：公司能够为下游客户及生态伙伴提供从0.2TOPs到6TOPs的不同算力水平的AIoT芯片，其中RK3588、RK3576带有6TOPs NPU处理单元，能够支持端侧主流的0.5B-3B参数级别的模型部署，可通过大语言模型实现翻译、总结、问答等功能，并可实现多模态搜索、识别，有效解决不同AIoT场景的痛点，提升产品使用体验。当前已有多个领域的客户基于瑞芯微主控芯片研发在端侧支持AI大模型的新硬件，例如教育平板、AI玩具、桌面机器人、算力终端、会议主机等产品。随着AI大模型在教育、家庭、医疗、工业、农业、服务业等边缘、端侧场景中持续加速落地，未来将赋能更多样的边、端侧AIoT产品。

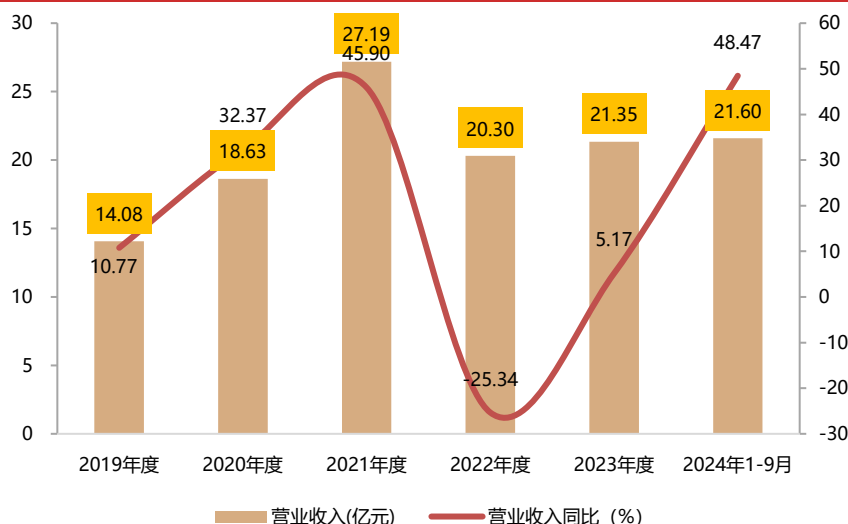
图表39：瑞芯微NPU介绍

SoC系列	应用	特点
RK1808	集成了初代RKNPU，相比传统CPU和GPU，在深度学习运算能力上有显著提升	3 TOPs for INT8 / 300 GOPs for INT16 / 100 GFLOPs for FP16；支持OpenCL/OpenVX；支持INT8/INT16/FP16；支持TensorFlow、Caffe、ONNX、Darknet模型
RK3399Pro		NPU up to 3.0TOPS，支持8bit/16bit运算，支持TensorFlow、Caffe模型
RV1109	第二代NPU，提升了NPU的利用率，适用于机器视觉应用	1.2Tops NPU，支持INT8/ INT16
RV1126		2.0Tops NPU，支持INT8/ INT16
RK3566	第三代NPU，搭载全新自研架构，适用于通用应用处理	采用四核64位Cortex-A55 CPU架构，搭载ARM G52 2EE GPU，1TOPS NPU
RK3568		采用四核64位Cortex-A55 CPU架构，主频最高2.0GHz，搭载ARM G52 2EE GPU，1TOPS NPU
RK3588	第四代NPU，自研架构再升级，支持多种AI应用	采用4核Cortex-A76 + 4核Cortex-A55的CPU架构，搭载ARM Mali-G610 MC4 GPU，并带6TOPs自研NPU，接口丰富度高，易于扩展；通常应用在目标方向的产品线中最高端的产品需求
RK3576	典型应用方向包括iFPD、工业控制及网关、云终端、人脸识别设备、车载中控、商显	采用4核Cortex-A72 + 4核Cortex-A53的CPU架构，搭载ARM Mali-G52 MC3GPU，带6TOPs自研NPU，满足各类人工智能应用，接口丰富度略小于RK3588；适用于中高端更主流档位的产品需求

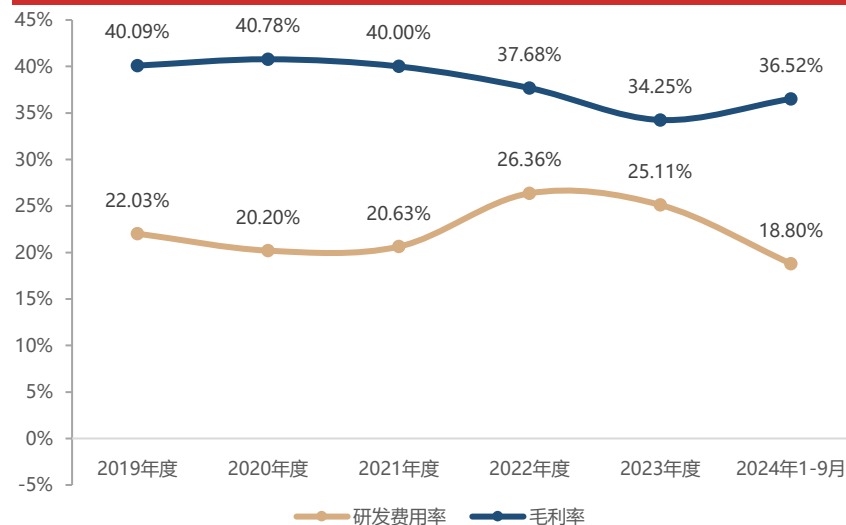
瑞芯微：持续推进NPU助力端侧AI

- **持续迭代自研核心IP：**公司坚持“IP芯片化”发展，基于“大音频、大视频、大感知、大软件”的核心技术方向，持续更新迭代自研的NPU、ISP、高清视频编解码、视频输出处理、视频后处理等核心IP。在人工智能领域，公司NPU IP从2018年至今历经多次迭代，对神经网络模型的支持和计算单元的利用效率不断提升，能够高效支持各种主流边端模型的本地化部署，赋能多场景边缘侧、端侧的AIoT产品应用。例如RK3588、RK3576采用高性能CPU和GPU内核并带有6T NPU处理单元，针对端侧主流的2B参数数量级别的模型运行速度能达到每秒生成10 token以上，满足小模型在边、端侧部署的需求。
- **各AIoT算力平台快速增长：**2024年公司实现营业收入约31~31.5亿元，创历史新高；实现净利润约5.5~6.3亿元，同比增长约307.75%~367.06%。第四季度公司延续了此前的成长逻辑，依托AIoT芯片“雁形方阵”布局优势，在旗舰芯片RK3588带领下，以多层次、满足不同需求的产品组合拳，促进公司长期深耕的AIoT多产品线占有率持续提升，尤其是在汽车电子、机器视觉、工业及行业类应用等领域；以RK3588，RK356X，RV11系列为代表的各AIoT算力平台快速增长。

图表40：2019-2024年1-9月瑞芯微收入情况



图表41：2019-2024年1-9月瑞芯微毛利率、研发费用率



资料来源：iFind，瑞芯微公告，中邮证券研究所

请参阅附注免责声明

更多免费行业报告

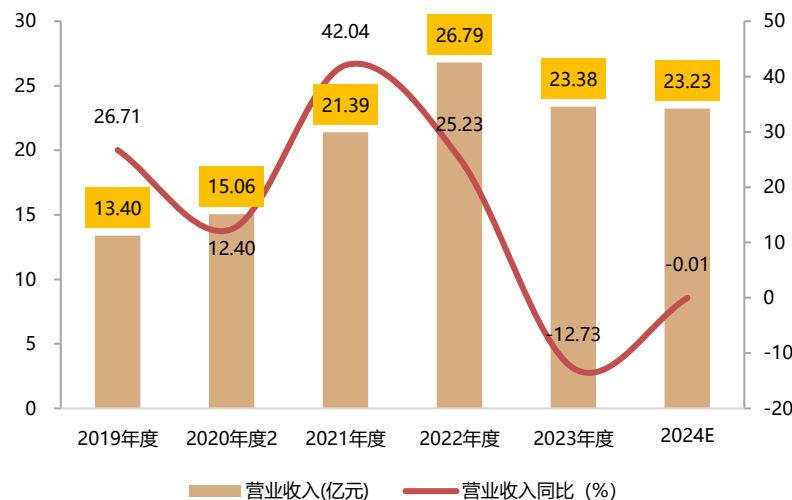
并购家

www.ipoipo.cn

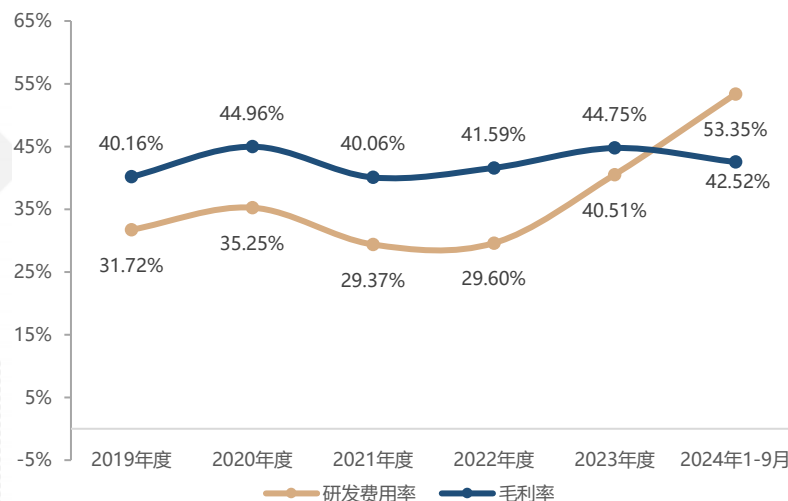
芯原股份：NPU IP赋能端侧AI

- 基于自有的IP，公司已拥有丰富的面向AI应用的软硬件芯片定制平台解决方案，涵盖如智能手表、AR/VR眼镜等实时在线（Always on）的轻量化空间计算设备，AI PC、AI手机、智慧汽车、机器人等高效率端侧计算设备，以及数据中心/服务器等高性能云侧计算设备。在端侧，公司积极布局智慧汽车、AR/VR等增量市场，已经为多家国际行业巨头客户提供了技术和服务。目前，集成了芯原NPU IP的人工智能（AI）类芯片已在全球范围内出货超过1亿颗，主要应用于物联网、可穿戴设备、智慧电视、智慧家居、安防监控、服务器、汽车电子、智能手机、平板电脑、智慧医疗等10个市场领域，在嵌入式AI/NPU领域全球领先，芯原的NPU IP已被82家客户用于上述市场领域的142款AI芯片中；在AR/VR眼镜领域，公司已为某知名国际互联网企业提供AR眼镜的芯片一站式定制服务，此外还有数家全球领先的AR/VR客户正在进行合作。

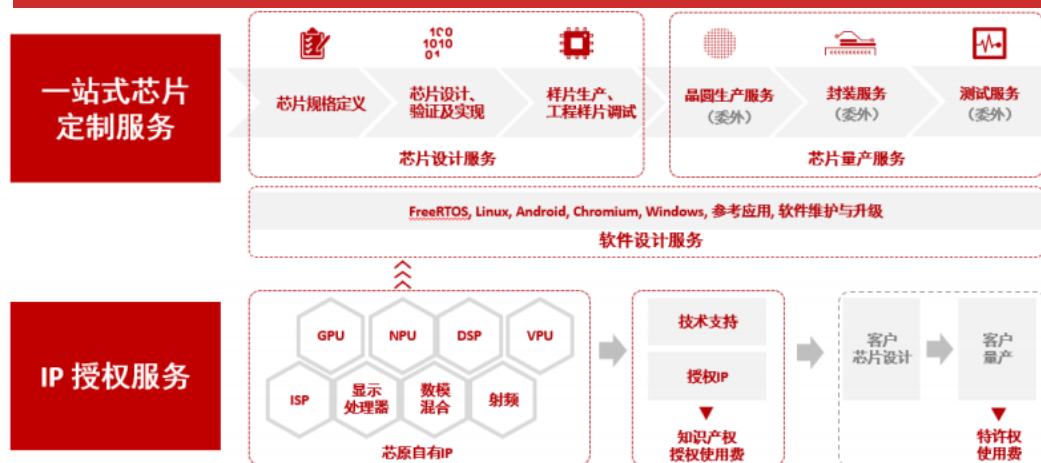
图表43：2019-2024E芯原收入情况



图表44：2019-2024年1-9月芯原毛利率、研发费用率



图表42：芯原股份产品



- 寒武纪是智能芯片领域全球知名的新兴公司，能提供云边端一体、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。公司掌握的智能处理器指令集、智能处理器微架构、智能芯片编程语言、智能芯片数学库等核心技术，具有壁垒高、研发难、应用广等特点，对集成电路行业与人工智能产业具有重要的技术价值、经济价值和生态价值。
- **AI赋能产业升级。**目前，公司已推出的产品体系覆盖了云端、边缘端的智能芯片及其加速卡、训练整机、处理器IP及软件，可满足云、边、端不同规模的人工智能计算需求。公司的智能芯片和处理器产品可高效支持视觉（图像和视频的智能处理）、语音处理（语音识别与合成）、自然语言处理以及推荐系统等多样化的人工智能任务，高效支持视觉、语音和自然语言处理等技术相互协作融合的多模态人工智能任务，广泛服务于大模型算法公司、服务器厂商、人工智能应用公司，可辐射智慧互联网、智能制造、智能教育、智慧金融、智能家居、智慧医疗等“智能+”产业。

图表45：核心技术框架

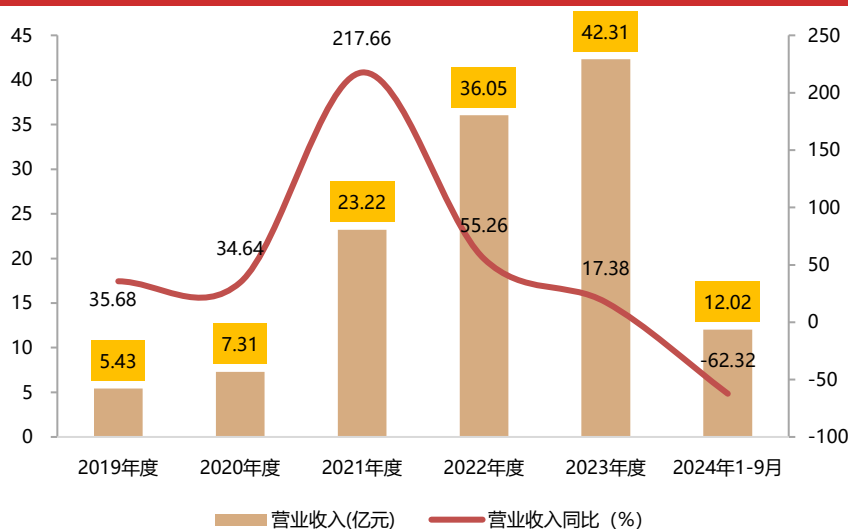


资料来源：寒武纪公告，中邮证券研究所

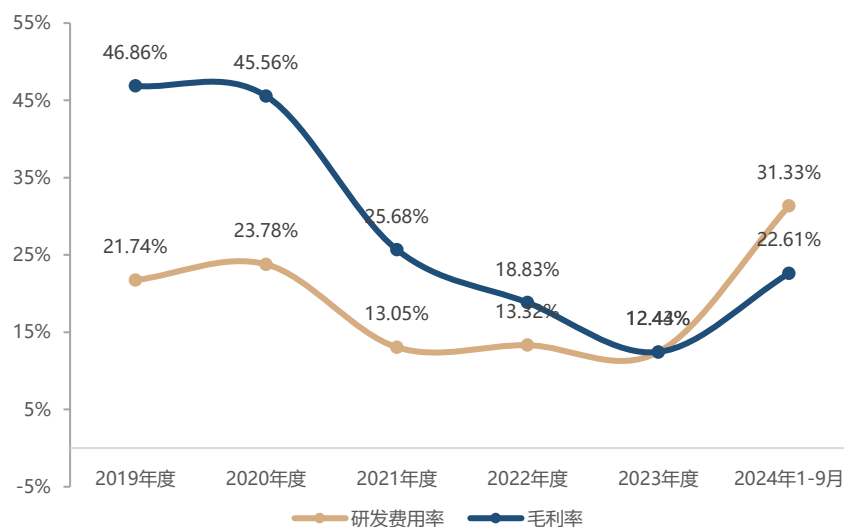
国科微：深度布局自研全场景NPU

- **深度布局自研全场景NPU，实现全系芯片标配NPU，为边端AI部署提供坚实的基座。**公司新一代4K AI视觉处理芯片GK7606V1系列搭载自研AI ISP引擎，内置双核A55处理器，拥有最高2.5TOPS算力，支持4K编解码、AI ISP、双3D降噪、图像防抖、多目拼接、多光谱融合等核心技术，具有高画质、低码率、低内存、低延时以及低功耗等优势，可用于智能安防、行车记录仪、无人机图传等领域，在客户端测试中表现抢眼，获得了国内主流安防企业与方案商的正向反馈。公司的AI ISP技术能够在极低照度的黑光场景下，对图像进行全方位的优化和增强，使图像更加清晰、细腻，色彩更加生动、饱和，演绎出“黑夜如白昼”的科技力量。公司新一代轻智能视觉处理芯片GK7203V1系列产品集成通用型轻算力NPU，支持双目输入与AOV，为经济型、消费级市场提供性价比更优的智能选择。公司智慧视觉类芯片产品的推出以及新品的发布将有助于公司产品价值及竞争力的提升，以此为契机，公司将深化与运营商、整机厂商以及方案商的合作，与产业链携手一起奔赴未来智能化的数字世界。未来，公司将进一步加强关键技术攻关，结合自研的NPU技术，优化边缘AI战略布局，加速AI能力在全系芯片的普及化应用。

图表46：2019-2024年1-9月国科微收入情况



图表47：2019-2024年1-9月国科微毛利率、研发费用率



资料来源：iFind，格隆汇，中邮证券研究所

请参阅附注免责声明

更多免费行业报告

并购家

www.ipoipo.cn

北京君正：持续进行NPU技术研发，不断丰富AI算法类型

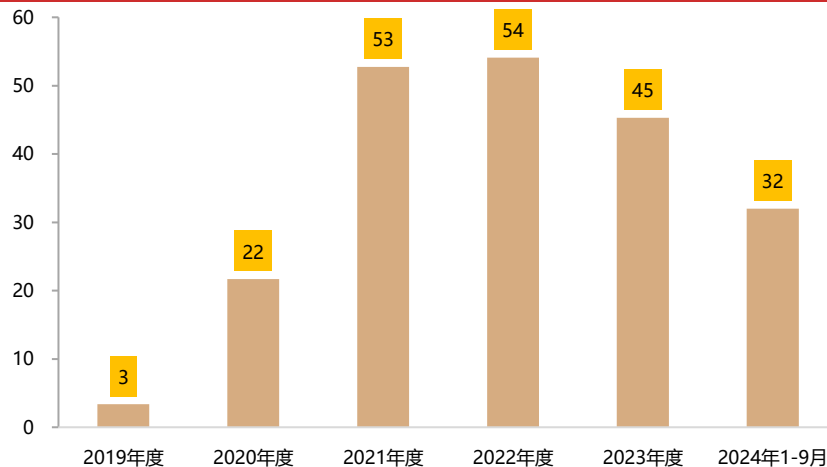
中邮证券
CHINA POST SECURITIES

- **持续进行NPU技术研发，不断丰富AI算法类型。**公司有算法、NPU等AI技术，应用在公司部分计算芯片产品中。公司的SOC芯片集成了公司自研的关键模块，包括CPU、VPU、NPU、ISP等，不同计算芯片产品有不同的配置。公司拥有自主研发的NPU技术已应用于T40、T41、A1等芯片中，且已量产销售。
- **安防IPC静待复苏。**从技术上讲，公司坚持核心IP自主研发并不断优化迭代比如VPU、ISP等，尤其是AI技术，有大量的算法已经有很多客户采用。公司的NPU自主研发，产品会根据不同的定位配置不同算力水平的NPU。从产品上讲，公司不断推出新品，优化产品布局，T31是前两年公司根据双摄的市场需求推出的一个产品，去年发挥了良好的市场推广作用，同时驱动公司亮眼的增长；T23是公司去年为今年规划的一个产品，今年还会推出T32，针对双摄需求的普及进一步优化了性能和性价比。目前公司在中低端的产品布局已经基本到位，未来中高端将是公司重要的发展方向之一。明年公司会推出T42的产品，C200也会逐步到位。
- **存储芯片静待拐点。**DRAM和Flash表现存在差异，三季度来看，DRAM环比二季度有一点下降，但降幅不大，Flash是增长的。DRAM市场中工业占比较大，所以工业市场需求疲软对DRAM的影响比较大，Flash方面，硬盘等消费类市场占比较多，所以和DRAM呈现不同的趋势，价格来看环比均有一点下降。

图表48：产品布局

X系列		T系列	
X2600		T41	
X2000		T40	
X1600		T31	
X1500	C系列	T23	A系列
X1000	C100	A1	

图表49：2019-2024年1-9月北京君正收入（亿元）



资料来源：北京君正官网，iFind，

请参阅附注免责声明

更多免费行业报告

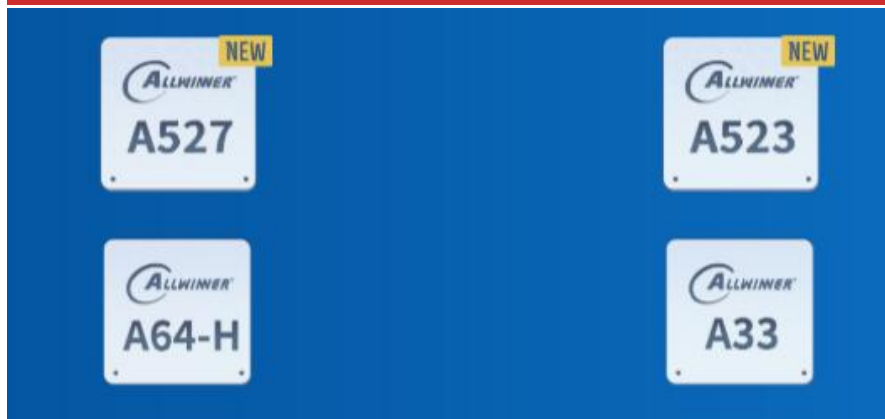
并购家

www.ipoipo.cn

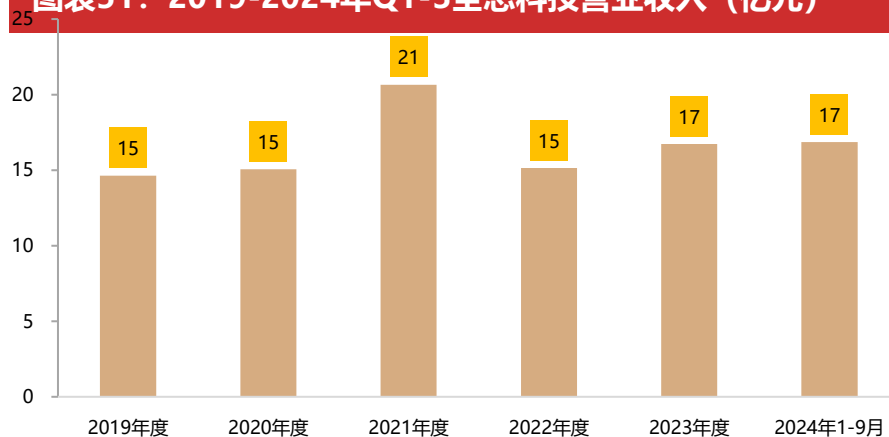
■ 紧密契合AI智能化需求，积极推动新品落地。

- 1) 在智能工业市场，公司推出基于八核AI机器人芯片MR527，MR527通过端侧CPU和NPU算力共同赋能，为视觉感知算法提供更强的AI端侧算力，进一步实现了更多障碍物的识别，有效改善了扫地机产品的避障能力，目前搭载该芯片的高端扫地机已对外发布并大规模量产。公司推出智能工业应用的T536和面向视觉AI扫地机机器人的MR536，目前已完成行业头部客户的送样；另外，相关产品均集成3T的NPU算力，满足AI场景的算力要求，加速工业和机器人的AI智能化；
- 2) 在智能汽车电子市场，搭载公司芯片的AR-HUD和智能激光大灯模块已在国内头部车企大规模量产，公司还推出了基于车规级八核异构通用计算平台T527V的产品方案以满足智能车载娱乐系统、全数字仪表等智能化模块应用需求；
- 3) 在智能终端领域，公司紧跟安卓最新生态的升级迭代，推出A523/A527系列高性能八核架构计算平台，公司上半年积极推广，相关产品已稳定量产，并获得了海内外众多终端平板品牌的认可和青睐。此外，公司基于智慧屏芯片H713系列，针对单片LCD光机特点进行深度优化和调校，提升了智能投影产品的画质体验，获得终端消费者高度认可，并成为主流的智能投影主控芯片供应商。

图表50：部分产品情况



图表51：2019-2024年Q1-3全志科技营业收入（亿元）



资料来源：全志科技官网，全志科技招股说明书，中邮证券研究所整理

请参阅附注免责声明

更多免费行业报告

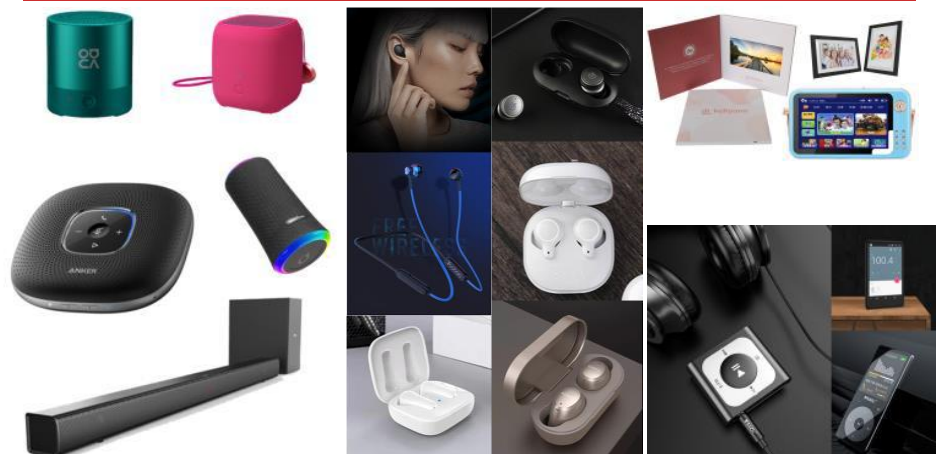
并购家

www.ipoipo.cn

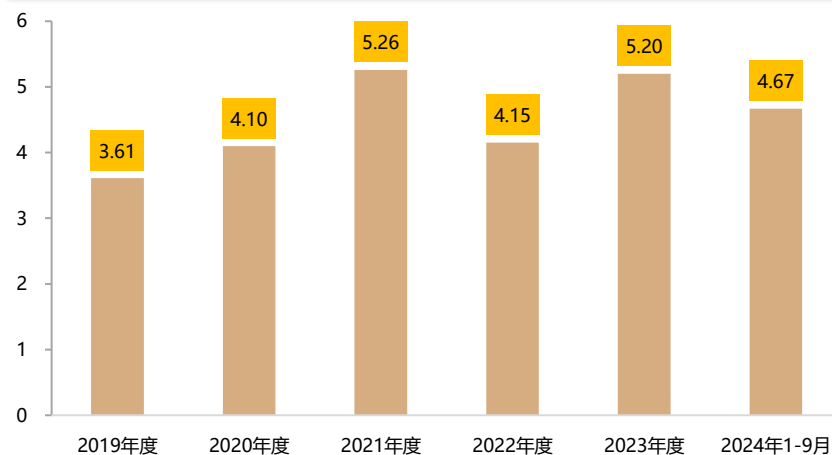
炬芯科技：三核AI异构芯片赋能端侧AI

- **存算一体积极拓展新市场。** 公司持续关注市场对低功耗、高算力端侧设备的需求，在家用音频领域投入新产品。存算一体AI处理器已导入多家客户，主要应用于高频音频降噪和人声处理，在音频降噪之外，扩展至手表等其他AI应用。公司已正式发布最新一代基于SRAM的模数混合存内计算的端侧AI音频芯片，采用CPU+DSP+NPU三核异构架构，可在更低功耗下提供更高算力，同时兼具更低的延迟和增强的安全性，将在音频应用和端侧AI中发挥重要作用。产品共包括三个芯片系列：第一个系列是ATS323X，面向低延迟私有无线音频领域；第二个系列是ATS286X，面向蓝牙AI音频领域；第三个系列是ATS362X，面向AI DSP处理器领域。目前部分客户已接近终端产品量产阶段。公司基于三核AI异构的芯片采用了更加先进的工艺制程，相较公司现有产品可以在现有功耗水平下提供几十倍至上百倍的算力提升，而相较于市场上主流的NPU产品能效比可以提升至少三倍以上，相较于主流的DSP产品在功耗方面能降低接近90%。基于SRAM的模数混合存内计算技术路径下的端侧AI音频芯片平台主要可以覆盖语音与音频、视觉识别以及健康类监测等相关应用场景，并且可实现端侧AI解决方案的快速落地。公司也将积极打造AI开发生态，借助炬芯完整工具链轻松实现算法的融合，帮助客户迅速地完成产品落地，助力AIoT产品AI化的不断演进。

图表52：公司产品

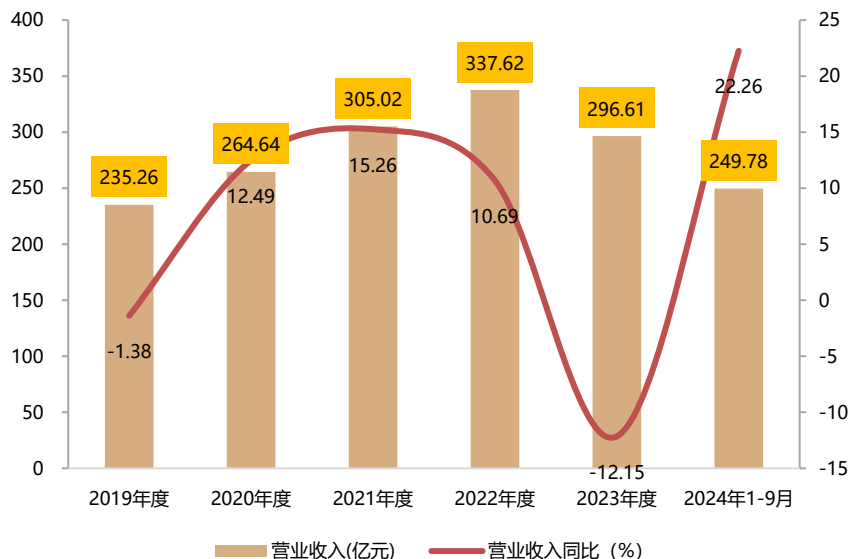


图表53：2019-2024年1-9月公司收入情况（亿元）

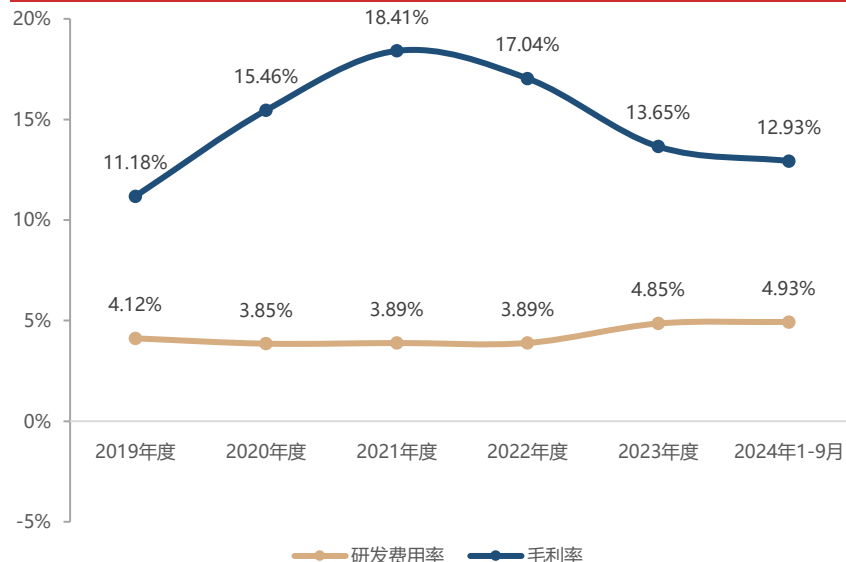


- **聚焦关键应用领域，面向全球市场，提供高端定制化封装测试解决方案和配套产能。**公司聚焦关键应用领域，在高算力及对应存储和连接、AI 端侧、功率与能源、汽车和工业等重要领域拥有行业领先的半导体先进封装技术（如 SiP、WL-CSP、FC、eWLB、PiP、PoP 及 XDFOI®系列等）以及混合信号/射频集成电路测试和资源优势，并实现规模量产，能够为市场和 客户提供量身定制的技术解决方案。在高性能先进封装领域，公司推出的 XDFOI® Chiplet 高密度多维异构集成系列工艺已按计划进入稳定量产阶段。该技术是一种面向Chiplet的极高密度、多扇出型封装高密度异构集成解决方案，其利用协同设计理念实现了芯片成品集成与测试一体化，涵盖2D、2.5D、3D集成技术， 公司正持续推进其多样化方案的研发及生产。经过持续研发与客户产品验证，公司XDFOI®不断取得突破，已在高性能计算、人工智能、5G、汽车电子等领域应用，为客户提供了外型更轻薄、数据传输速率更快、功率损耗更小的芯片成品制造解决方案，满足日益增长的终端市场需求。

图表54：2019-2024年1-9月长电科技收入情况



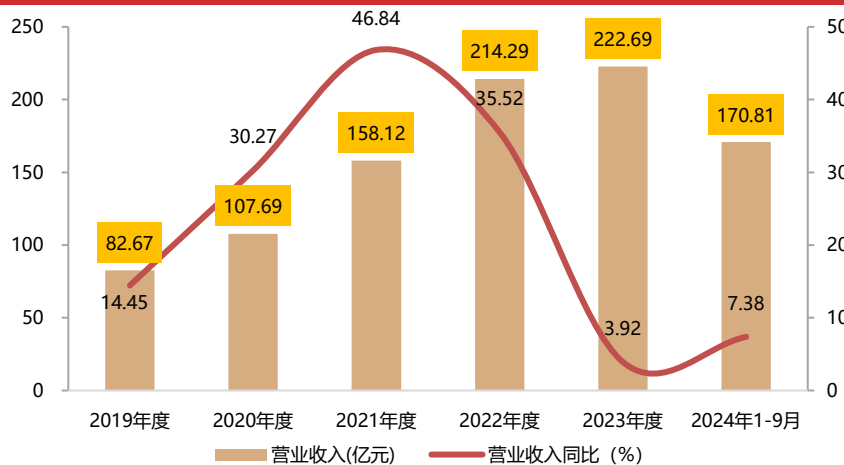
图表55：2019-2024年1-9月长电科技毛利率、研发费用率



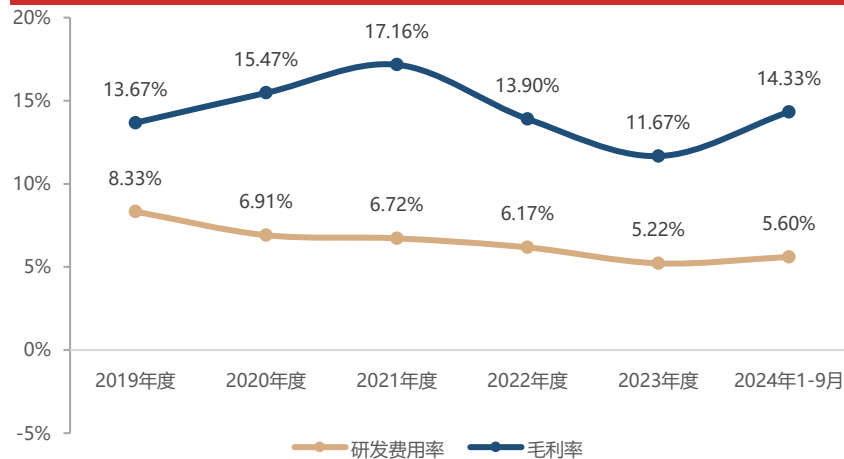
资料来源：iFind，长电科技公告，中邮证券研究所

- **拟间接持有引线框架供应商AAMI股权加速先进封装布局。**2024年10月16日，公司与领先半导体签署了《合伙份额转让协议》，公司拟出资2亿元受让领先半导体持有的滁州广泰146,722,355.58元出资额（占滁州广泰合伙份额的31.90%），以间接持有引线框架供应商AAMI股权。AAMI专业从事引线框架的设计、研发、生产与销售。引线框架借助于键合材料实现芯片内部电路引出端与外引线的电气连接，形成电气回路，起到和外部导线连接的桥梁作用，在大部分半导体产品中均有应用。AAMI在引线框架领域深耕超过40年，是全球前列、国内领先的引线框架供应商，拥有先进的生产工艺、高超的技术水平和强大的研发能力，积累了丰富的产品版图、技术储备和客户资源，在高精密和高可靠性等高端应用市场拥有极强的竞争优势，产品广泛应用于汽车、计算、工业、通信及消费类半导体，得到各细分领域头部客户的高度认可，与全球顶尖的半导体IDM和封测代工企业建立了稳固的合作关系。
- **收购京隆科技26%股权丰富高端集成电路专业测试布局。**2025年2月13日，公司披露关于收购京隆科技（苏州）有限公司26%股权交割完成的公告，本次交割完成后，公司持有京隆科技26%的股权。京隆科技运营模式和财务状况良好，其在高端集成电路专业测试领域具备差异化竞争优势。

图表56：2019-2024年1-9月通富微电收入情况



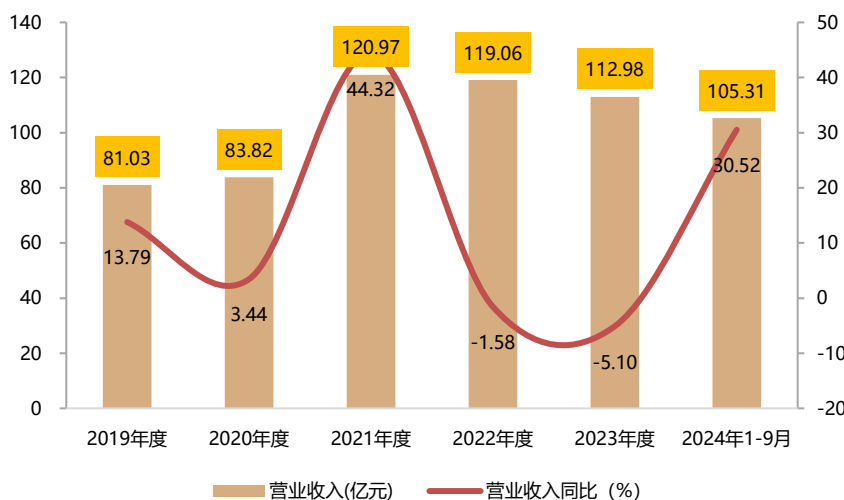
图表57：2019-2024年1-9月通富微电毛利率、研发费用率



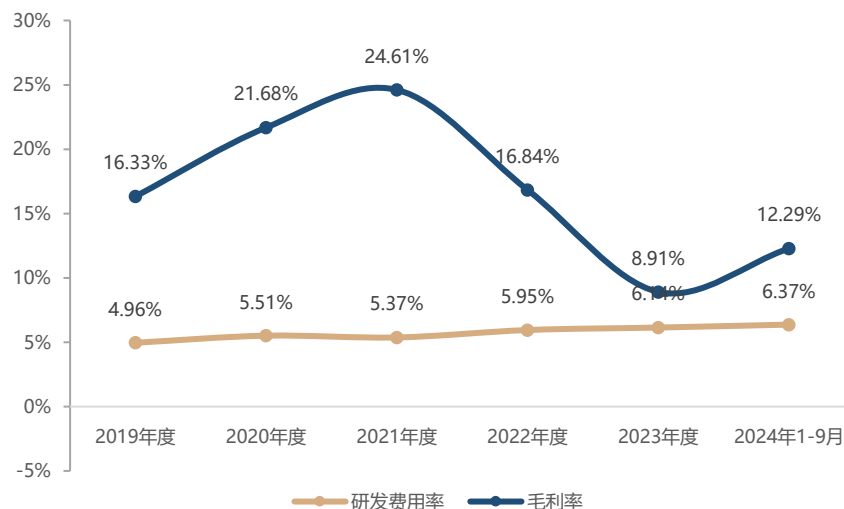
资料来源：iFind，通富微电公告，中邮证券研究所

■ **重点研发Fan-Out、FOPLP、汽车电子、存储器等先进封装技术和封装产品。**公司的主要生产基地有天水、西安、昆山、南京、韶关、Unisem以及刚投产的江苏和上海。天水基地以引线框架类产品为主，产品主要涉及驱动电路、电源管理、蓝牙、MCU、NOR Flash等。西安基地以基板类和QFN、DFN产品为主，产品主要涉及射频、MEMS、指纹产品、汽车电子、MCU、电源管理等。南京基地以存储器、射频、MEMS等集成电路产品的封装测试为主。昆山基地封装晶圆级产品，主要产品包括TSV、Bumping、WLCSP、Fan-Out等。韶关基地以引线框架类封装产品、显示器件和显示模组产品为主。Unisem封装产品包括引线框架类、基板类以及晶圆级产品，主要以射频类产品为主。华天科技（江苏）有限公司、上海华天集成电路有限公司2024年刚投产，华天江苏封装的产品有Bumping、WLCSP、Fan-Out等晶圆级产品，华天上海主要开展晶圆测试和成品测试业务。2024年度，在相关电子终端产品需求回暖的影响下，集成电路景气度回升。受此影响，报告期内，公司订单增加，产能利用率提高，营业收入较去年同期有显著增长，从而使得公司经营业绩大幅提高。公司2024年预计实现归母净利润5.5-6.3亿元，同比增长143.02% - 178.36%。

图表58：2019-2024年1-9月华天科技收入情况



图表59：2019-2024年1-9月华天科技毛利率、研发费用率



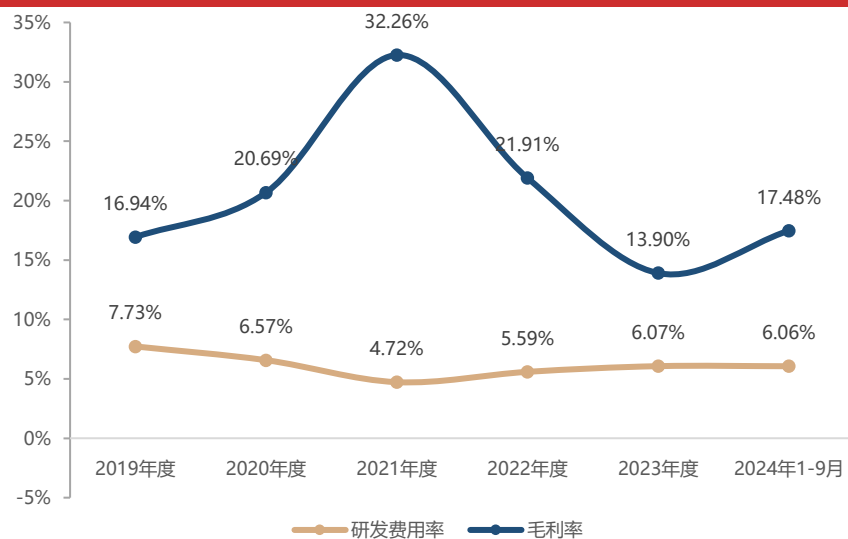
资料来源：iFind，华天科技公告，中邮证券研究所

- **AI助力SOC客户持续成长。**在客户端，公司已成为国内众多SoC类客户的第一供应商，伴随客户一同成长，另外公司在台系客户方面不断推进，持续贡献营收。公司坚持服务于中高端客户，聚焦先进封装领域，坚持做有门槛的客户及有门槛的产品。公司目前已经形成以细分领域龙头设计公司为核心的客户群，公司现有以及潜在客户可以分成三类：现有的以大陆地区高成长SoC类设计公司为代表的核心客户群，公司作为这些客户的核心供应商，通过不断承接其新产品和提升市场份额，伴随其一同成长；海外特别是中国台湾地区的头部客户；除此之外，其他海外客户如欧美客户和国内HPC、汽车电子领域的客户群的增长。
- **盈利能力逐步释放。**公司处于高速成长阶段，近两年整体投资规模较大，短期内确实对利润产生一定影响。从财务角度看，随着公司营收规模的增长，规模效应显现，公司毛利率会逐步上升，费用率会下降。公司目前管理费用率和财务费用率较高，主要是因为二期项目人才招募多、贷款规模大，随着经营规模的扩大和融资渠道多元化，后续费用率会持续下降。得益于良好的客户结构和产品结构，公司目前整体毛利率仍然位于行业前列，相信随着产能爬坡和规模效应的逐渐体现，未来利润端也会逐渐体现。

图表60：2019-2024年甬矽电子收入情况

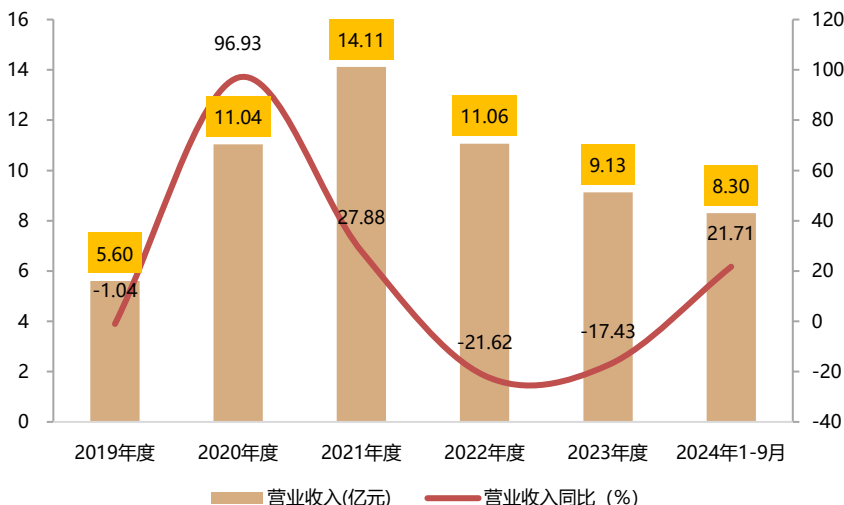


图表61：2019-2024年1-9月甬矽电子毛利率、研发费用率

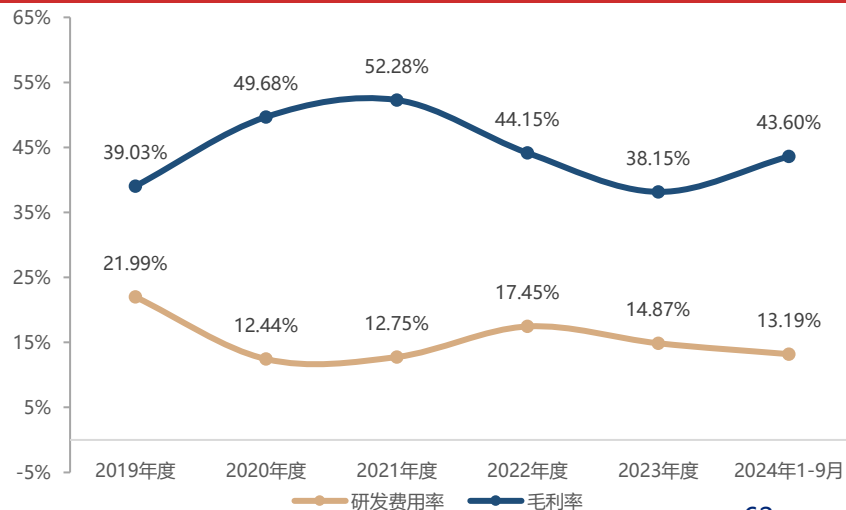


- **海外产能积极布局。**依托新加坡子公司国际业务总部，进一步完善公司海外业务中心、研发工程中心与投融资平台，并积极推进马来西亚槟城生产与制造基地的筹备与建设，以更好贴近海外客户需求、巩固提升海外产业链地位，推进工艺创新与项目开发，搭建布局全球化的投融资平台与生产制造基地。
- **新技术不断布局。**公司作为晶圆级硅通孔（TSV）封装技术的领先者，聚焦以影像传感芯片为代表的智能传感器市场，持续根据产品与市场新需求，对工艺进行创新优化。不断提升车规STACK封装技术的工艺水平与量产能力，推进A-CSP工艺的开发拓展，扩大在车载CIS领域的技术领先优势与生产规模；发挥产能、技术、核心客户等优势，进一步巩固在智能手机、安防监控数码等应用领域市场占有率；积极布局拓展新的应用市场，大力推进MEMS、Filter、AR/VR等应用领域的商业化应用规模；积极开发晶圆级集成封装技术，把握产业和市场不断变化的创新需求。加强微型光学器件设计、研发与制造能力的整合与拓展。发挥Anteryon公司领先的光学设计与开发能力，通过设备投资与团队引进，持续提升混合镜头产品在半导体设备、智能制造、农业自动化市场的业务规模；积极扩大晶圆级光学器件（WLO）制造技术在汽车智能投射领域的量产规模，并加大与TIE1的共同合作，努力推进汽车大灯、信号灯等车用智能交互系统产品的开发进程。

图表62：2019-2024年1-9月晶方科技收入情况



图表63：2019-2024年1-9月晶方科技毛利率、研发费用率



- AI端侧发展不及预期风险。

感谢您的信任与支持!

THANK YOU

吴文吉 (首席分析师)

SAC编号: S1340523050004

邮箱: wuwenji@cnpsec.com

翟一梦 (研究助理)

SAC编号: S1340123040020

邮箱: zhaiyimeng@cnpsec.com

分析师声明

撰写此报告的分析师（一人或多人）承诺本机构、本人以及财产利害关系人与所评价或推荐的证券无利害关系。

本报告所采用的数据均来自我们认为可靠的目前已公开的信息，并通过独立判断并得出结论，力求独立、客观、公平，报告结论不受本公司其他部门和人员以及证券发行人、上市公司、基金公司、证券资产管理公司、特定客户等利益相关方的干涉和影响，特此声明。

免责声明

中邮证券有限责任公司（以下简称“中邮证券”）具备经中国证监会批准的开展证券投资咨询业务的资格。

本报告信息均来源于公开资料或者我们认为可靠的资料，我们力求但不保证这些信息的准确性和完整性。报告内容仅供参考，报告中的信息或所表达观点不构成所涉证券买卖的出价或询价，中邮证券不对因使用本报告的内容而导致的损失承担任何责任。客户不应以本报告取代其独立判断或仅根据本报告做出决策。

中邮证券可发出其它与本报告所载信息不一致或有不同结论的报告。报告所载资料、意见及推测仅反映研究人员于发出本报告当日的判断，可随时更改且不予通告。

中邮证券及其所属关联机构可能会持有报告中提到的公司所发行的证券头寸并进行交易，也可能为这些公司提供或者计划提供投资银行、财务顾问或者其他金融产品等相关服务。

《证券期货投资者适当性管理办法》于2017年7月1日起正式实施，中邮证券客户中的专业投资者使用，若您非中邮证券客户中的专业投资者，为控制投资风险，请取消接收、订阅或使用本报告中的任何信息。本公司不会因接收人收到、阅读或关注本报告中的内容而视其为专业投资者。

本报告版权归中邮证券所有，未经书面许可，任何机构或个人不得存在对本报告以任何形式进行翻版、修改、节选、复制、发布，或对本报告进行改编、汇编等侵犯知识产权的行为，亦不得存在其他有损中邮证券商业性权益的任何情形。如经中邮证券授权后引用发布，需注明出处为中邮证券研究所，且不得对本报告进行有悖原意的引用、删节或修改。

中邮证券对于本申明具有最终解释权。

公司简介

中邮证券有限责任公司，2002年9月经中国证券监督管理委员会批准设立，注册资本50.6亿元人民币。中邮证券是中国邮政集团有限公司绝对控股的证券类金融子公司。

公司经营范围包括：证券经纪；证券自营；证券投资咨询；证券资产管理；融资融券；证券投资基金销售；证券承销与保荐；代理销售金融产品；与证券交易、证券投资活动有关的财务顾问。此外，公司还具有：证券经纪人业务资格；企业债券主承销资格；沪港通；深港通；利率互换；投资管理人受托管理保险资金；全国银行间同业拆借；作为主办券商在全国中小企业股份转让系统从事经纪、做市、推荐业务资格等业务资格。

公司目前已经在北京、陕西、深圳、山东、江苏、四川、江西、湖北、湖南、福建、辽宁、吉林、黑龙江、广东、浙江、贵州、新疆、河南、山西、上海、云南、内蒙古、重庆、天津、河北等地设有分支机构，全国多家分支机构正在建设中。

中邮证券紧紧依托中国邮政集团有限公司雄厚的实力，坚持诚信经营，践行普惠服务，为社会大众提供全方位专业化的证券投、融资服务，帮助客户实现价值增长，努力成为客户认同、社会尊重、股东满意、员工自豪的优秀企业。

投资评级说明

投资评级标准	类型	评级	说明
报告中投资建议的评级标准： 报告发布日后的6个月内的相对市场表现，即报告发布日后的6个月内的公司股价（或行业指数、可转债价格）的涨跌幅相对同期相关证券市场基准指数的涨跌幅。 市场基准指数的选取：A股市场以沪深300指数为基准；新三板市场以三板成指为基准；可转债市场以中信标普可转债指数为基准；香港市场以恒生指数为基准；美国市场以标普500或纳斯达克综合指数为基准。	股票评级	买入	预期个股相对同期基准指数涨幅在20%以上
		增持	预期个股相对同期基准指数涨幅在10%与20%之间
		中性	预期个股相对同期基准指数涨幅在-10%与10%之间
		回避	预期个股相对同期基准指数涨幅在-10%以下
	行业评级	强于大市	预期行业相对同期基准指数涨幅在10%以上
		中性	预期行业相对同期基准指数涨幅在-10%与10%之间
		弱于大市	预期行业相对同期基准指数涨幅在-10%以下
	可转债评级	推荐	预期可转债相对同期基准指数涨幅在10%以上
		谨慎推荐	预期可转债相对同期基准指数涨幅在5%与10%之间
		中性	预期可转债相对同期基准指数涨幅在-5%与5%之间
		回避	预期可转债相对同期基准指数涨幅在-5%以下

中邮证券研究所

北京

邮箱：yanjiusuo@cnpsec.com

地址：北京市东城区前门街道珠市口东大街17号

邮编：100050

上海

邮箱：yanjiusuo@cnpsec.com

地址：上海市虹口区东大名路1080号大厦3楼

邮编：200000

深圳

邮箱：yanjiusuo@cnpsec.com

地址：深圳市福田区滨河大道9023号国通大厦二楼

邮编：518048

