电子行业

评级：增持（首次）

2025 年 2 月 27 日

AI 应用侧深度渗透，驱动国产先进封装技术寻求突破

行业指数相对沪深 300 表现

证券分析师：唐仁杰
执业证书编号：S0370524080002
公司邮箱：tangrj@jyzq.cn
联系电话：0755-83025184

- DeepSeek 在算法层面实现三大突破——通过低秩键值压缩（MLA）将注意力计算内存占用降低 80%，动态稀疏 MoE 架构使每个 Token 仅激活 5.5%参数，以及 GRPO 强化学习框架驱动模型自主进化多步推理能力。千亿参数模型在通用任务上达到与密集模型相当的精度，同时降低 37%推理延迟。**模型的高效运行仍依赖硬件层面的三重能力支撑：高并行计算、高存储带宽、超低延迟互连。**

- **效率提升≠需求下降：**本质上，算法优化并非削弱算力产业价值，而是通过重构需求结构打开更大市场空间——从集中式训练向分布式推理延展，从通用计算向场景专用架构升级，最终形成万亿级算力市场的多级增长引擎。**"降本→普及→增量"的螺旋上升效应将推动 Post-training 微调算力激增、云端推理并发量指数增长、边缘侧长尾需求爆发带来总算力需求。**

- **模型参数量、训练数据持续扩充，高性能算力芯需求仍高：**单纯倚仗传统芯片设计与制造通过缩小 FET 尺寸去提高芯片性能的方式效率降低，且规模化边际减弱。更重要的是，对于不同场景化需求不同，高带宽，低延迟，高能效比有更高要求，**系统级线宽/线距瓶颈限制了高速数据在芯片之间、芯片与外部存储器之间高效传输，严重制约了 AI 芯片性能的充分释放。先进封装是“More Than Moore”（超越摩尔）时代的解决方案。**

- **封装技术正逐步从 PCB 的层面，向芯片内部（即 IC 层面）转变：**采用 2.5D 和 3D 封装技术，不再依赖传统的 PCB 作为主连接平台，而是直接将多个 IC 芯片通过转接板（interposer，如硅转接板、玻璃转接板等）进行集成。**2.5D 封装技术的核心在于 TSV、Interposer、RDL、Bumps，各大厂商基于这些组装以达到不同客户需求。据 YOLE 预测，2023 年全球先进封装营收约 378 亿美元，占半导体封装市场的 44%；2024 年增长至 425 亿美元，至 2029 年，先进封装营收有望增长至 695 亿美元，年复合增长率 11%，其中 2.5D/3D 封装渗透率最快。**

- **投资建议：关注 2.5D/3D 封装技术核心前道设备厂商、基板材料及 OSAT 厂商。**
  - **设备厂商：北方华创、拓荆科技、盛美上海、中微公司**
  - **基板材料厂商：兴森科技**
  - **OSAT 厂：长电科技、通富微电**

- **风险提示：** 1、2.5D\3D 封装及其他先进封装难度较大，良率有待改善，或影响利润；2、前期设备投入及研发成本较高；3、AI 应用落地速度不及预期

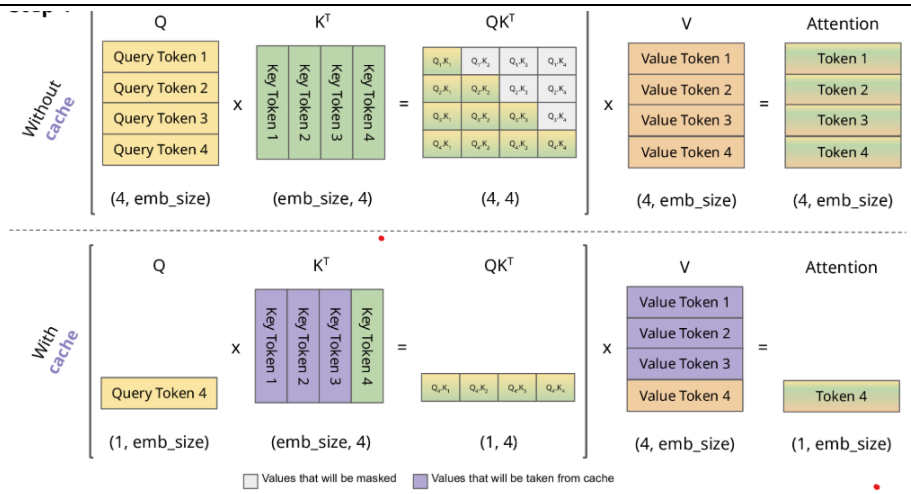
一、DeepSeek 架构上的突破-算法层面解决算力效率问题

DeepSeek 从模型的输入处理阶段到计算阶段再到模型的输出阶段进行深层次优化，显著提升算力效率，使得其在训练阶段以及推理阶段在保持模型性能的同时，减少冗余计算，从而塑造出更高性价比模型。

传统 Transformer 模型的自注意力机制存在显著的计算瓶颈：处理 n 长度序列时需构建 n²规模的注意力矩阵，导致内存和计算复杂度均呈$O(n^2)$增长。以 1024 长度序列为例，单头注意力矩阵即需 4MB 存储，叠加多头多层结构后硬件资源极易耗尽。在推理场景中，由于需实时逐 Token 生成文本，重复计算历史 Token 的键值数据会引发指数级资源消耗。

DeepSeek 通过引入 KV 缓存机制实现突破性优化：将历史 Token 的键值向量存储复用，仅计算新 Token 的查询向量进行匹配。该策略使推理阶段复杂度从$O(n^2)$降至$O(n)$，大幅减少冗余计算。**KV 缓存快速存取，以及更强的并行计算能力处理动态增长的序列数据，仍对高性能算力芯片吞吐量有一定要求。**

图表 1：With KV cache VS without KV cache

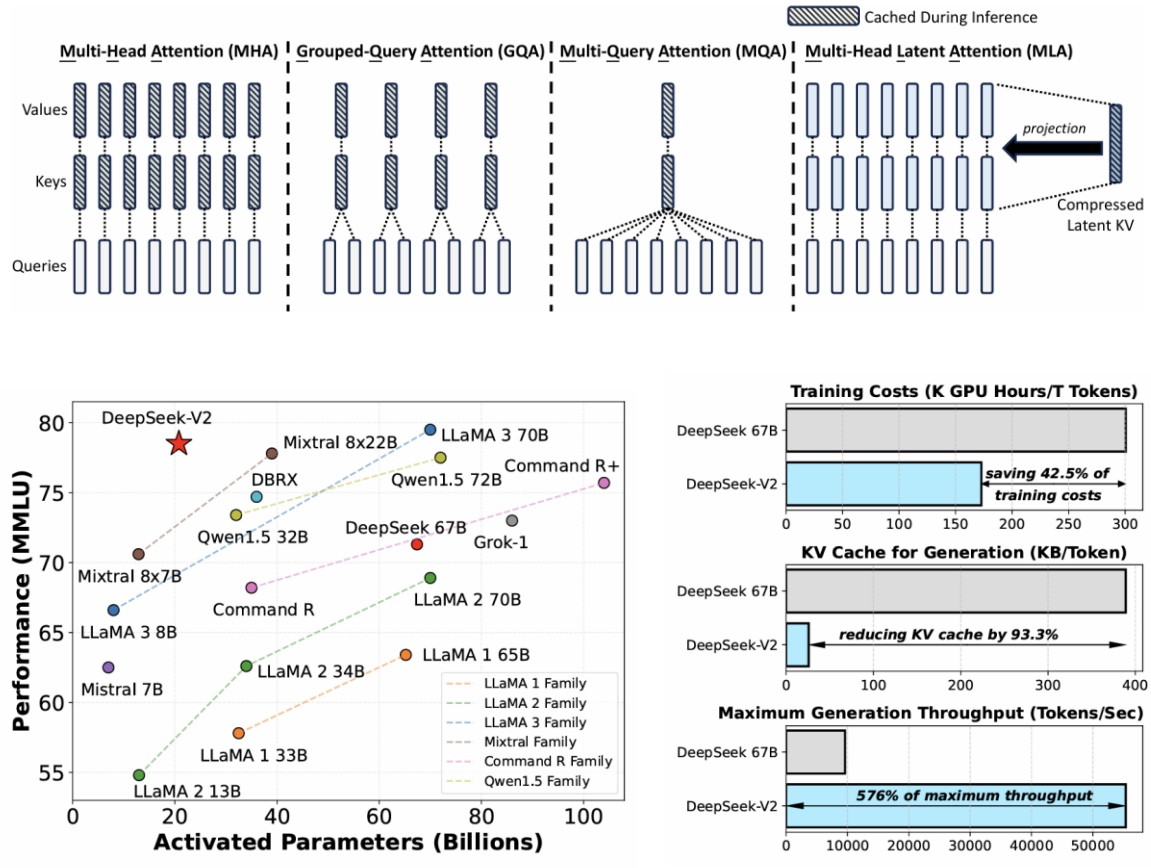


数据来源：Transformers KV Caching Explained，金元证券研究所

DeepSeek V2 通过 Multi-Head Latent Attention（MLA）技术突破现有注意力机制瓶颈：传统多头注意力（MHA）需存储完整键值矩阵，导致 KV 缓存空间随序列长度线性膨胀。主流改进方案如 MQA（多查询注意力）和 GQA（分组查询注意力）虽能降低缓存需求，但存在显著性能损失——MQA 缓存需求最小但精度最弱，GQA 则在缓存与性能间折中。

MLA 创新性地引入低秩键值联合压缩：将原始高维键值矩阵映射至低秩潜在空间，仅需存储压缩后的潜在向量。该方法使 KV 缓存空间较 MHA 减少 90%以上（对标 GQA 水平），同时保持与 MHA 相当的性能表现。

图表 2：MHA vs GQA vs MQA vs MLA



数据来源：DeepSeek V2 tech report，金元证券研究所

DeepSeek-V3 的混合专家（MoE）架构实现超大规模高效计算

相较于传统 Dense 模型（如 Llama3），DeepSeek-V3 作为 6710 亿参数的 MoE 模型，通过动态稀疏计算突破算力瓶颈：每个 Token 仅激活约 5.5%参数（37B/671B），在保持模型规模优势的同时显著降低计算负载：

- **动态路由机制：**通过门控网络为每个 Token 选择 1-2 个专家（小型前馈神经网络），替代传统 Transformer 中全参数参与的固定计算模式。

- **稀疏计算流：**仅被选中的专家执行正向传播，其余 90%以上参数处于静默状态。**通过细粒度专家+共享专家的组合替换粗粒度的专家，形成更高细粒度的专家池。**

训练效率方面：

- 正向传播：单步计算量较 Dense 模型减少 40%-60%（与专家选择数量强相关）
- 反向传播：梯度更新仅作用于被激活的专家及路由网络，参数更新量减少至全量模型的 10%以下

推理效率方面：

- 算力需求解耦：推理延迟与激活参数量（而非总参数量）正相关，长文本处理效率提升 3-5 倍
- 硬件友好性：稀疏计算模式更适配支持动态路由的 AI 加速芯片

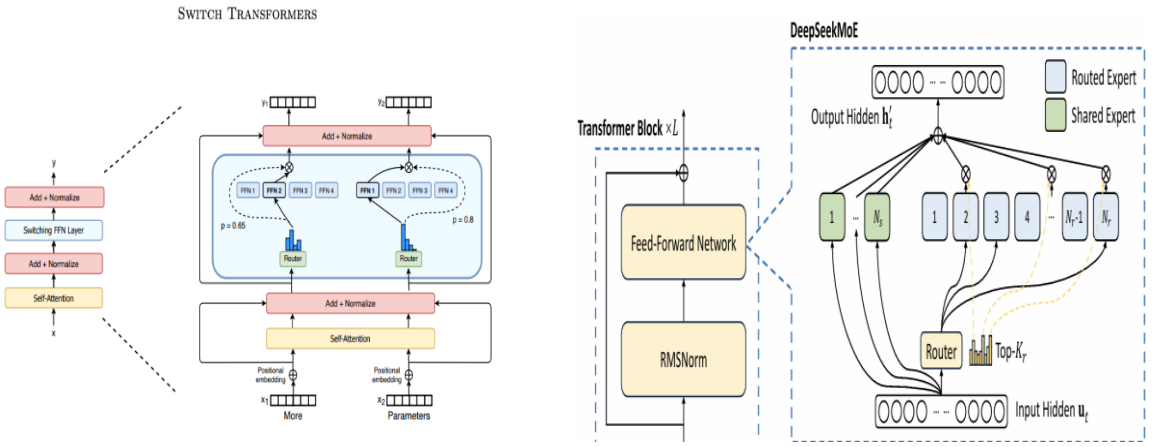
架构使模型在同等算力预算下，可扩展至 10 倍于 Dense 模型的参数量，为"规模决定性能"的大模型发展提供可持续路径。

图表 3：DeepSeek-V3 采用 DeepSeek MoE 架构，算法层面提升计算效率

	DeepSeek V3	DeepSeek V2.5 0905	Qwen2.5 72B-Inst	Llama3.1 405B-Inst
Architecture	MoE	MoE	DENSE	DENSE
# Activated Params	37B	21B	72B	405B
# Total Params	671B	236B	72B	405B

数据来源：DeepSeek，金元证券研究所

图表 4：MoE 基础架构



数据来源：Hugging Face，Switch Transformer by Google，DeepSeek V3 tech report，金元证券研究所

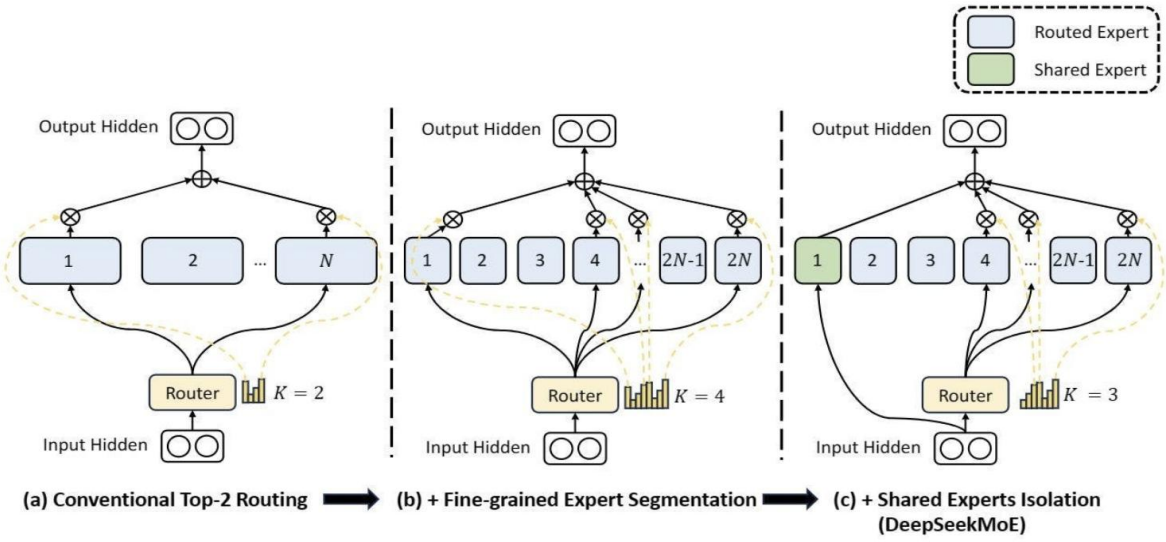
图 1 左侧：标准 Transformer 块；右侧：Switch Transformer 块，单个 FFN 替换为多个 FFN（名为“专家”）

传统 MoE 模型通过引入辅助损失函数强制均衡专家负载，但策略因忽视数据分布特性，导致同类任务被分散路由至不同专家，引发领域知识割裂与参数冗余两大问题。**DeepSeek V3 创新性提出无辅助损失负载均衡策略**，在门控网络中嵌入可学习偏置项，动态感知专家负载状态并自动调节路由偏好：过载专家通过偏置负向修正降低激活概率，使模型在训练过程中自主收敛至负载均衡

与知识聚合的平衡态。通过共享专家（每层强制全局激活）与 256 个路由专家的协同设计，**模型在 Token 级别动态筛选 8 个候选专家并最终路由至≤4 个高置信节点，实现通用能力集中化（共享专家承载跨领域知识）与专业能力垂直化（路由专家专注细分任务）的解耦优化。**

训练阶段容忍 10:1 的专家激活频率差异，通用任务精度损失<0.5%的同时提升垂直任务性能 12-15%；推理阶段通过共享专家固定激活与动态路由的混合计算流，单步计算量较传统 MoE 减少 37%，显存占用下降 28%。技术突破对算力芯片提出新需求，**即需支持偏置项实时更新（微秒级动态路由决策）与专家权重异构存储（共享专家高频访问数据独立缓存）**

图表 5： 共享专家+无额外损耗负载均衡策略



数据来源：DeepSeek V3，金元证券研究所

二、DeepSeek-R1：打造更强大推理能力

DeepSeek-R1 系列包含基础模型 R1-Zero 及其蒸馏变体，突破性地通过纯强化学习（RL）路径实现大语言模型高阶推理能力，颠覆"监督微调（SFT）为推理能力必要前提"的传统认知。其核心创新在于群体相对策略优化（GRPO）算法，相较主流近端策略优化（PPO）实现三大技术跃迁：

1、算法架构重构

- **去价值模型依赖：**GRPO 摒弃 PPO 中独立的价值模型（Value Model），通过组内相对优势计算替代绝对基线预测，消除策略-价值模型协同训练的开销
- **动态组评分机制：**对同批次生成结果进行组内排序，基于相对奖励积分（如 Top 20%答案自动获得优势权重）驱动策略更新，避免 PPO 中广义优势估计（GAE）的复杂计算

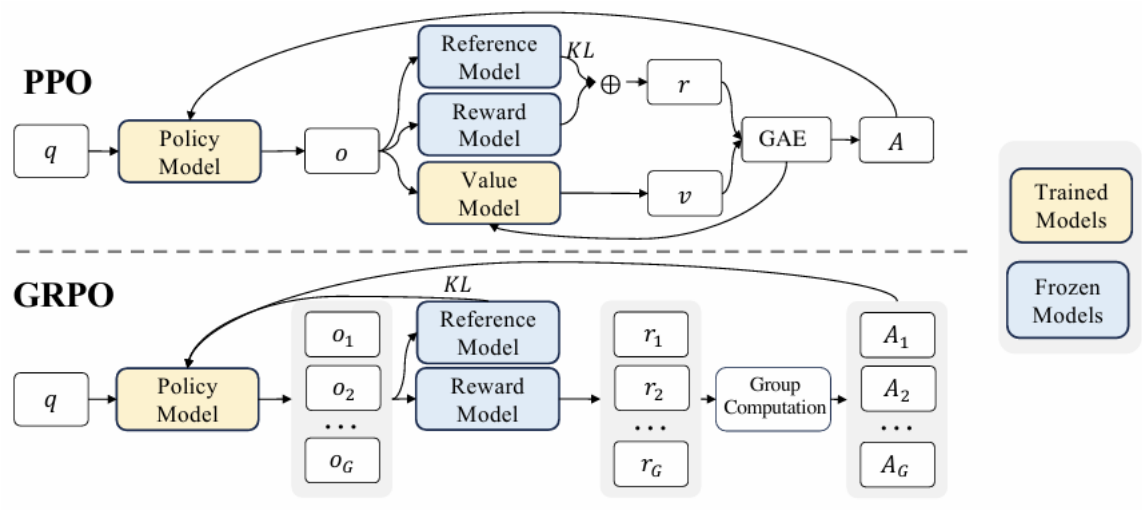
2、计算效率突破

- **训练成本对比：**在同等 7B 参数规模下，GRPO 较 PPO 减少 32%显存占用，单步训练耗时下降 41%
- **收敛效率提升：**在代码生成任务（HumanEval 基准）中，GRPO 达成 80%最终性能的迭代轮次仅为 PPO 的 1/3

3、 推理能力强化路径：

- 推理（Reasoning）与推断（Inference）解耦：
  - **推理能力：**通过 GRPO 的组内对抗机制，强制模型学习逻辑链拆解与多步决策优化（如数学证明题解决路径规划）
  - **推断能力：**保留基础 Transformer 架构的并行计算特性，确保 Token 生成速度与标准模型对齐

图表 6：GRPO vs PPO



数据来源：DeepSeek Math，金元证券研究所

在 GSM8K 数学推理数据集上，R1-Zero 未经过 SFT 直接通过 GRPO 训练，准确率达 82.3%，超越同规模 SFT+PPO 方案（78.1%）

DeepSeek-R1-Zero 通过群体相对策略优化（GRPO）算法，在纯强化学习框架下实现了大语言模型推理能力的自主进化，其核心突

破在于无需监督微调（SFT）即可完成高阶逻辑思维的涌现。训练过程中，模型展现出显著的非线性能力跃迁：初期阶段（0-30%训练周期）的思考链长度局限在 50-100 Token，仅能处理简单推理任务（如基础算术），但在引入 GRPO 的组内相对优势机制后，模型自发扩展多步推理能力，复杂数学证明任务的思考链长度提升至 2000+ Token，且伴随策略梯度突变现象（训练损失曲率变化超 40%），驱动 AIME-2024 评测的 pass@1 准确率从 15.6%跃升至 71.0%，达到与 SFT+RLHF 混合训练方案相当的水平。

这一过程依赖于 GRPO 构建的自监督探索-评估闭环——模型通过批量生成候选答案并动态对比组内奖励积分，自主优化推理路径规划策略，例如在 33%训练周期后出现关键转折点：错误答案的路径回溯率提升 62%，高难度任务的计算资源占比从 15%增至 58%，实现类似人类"顿悟"(Aha Moment)的策略优化效果。

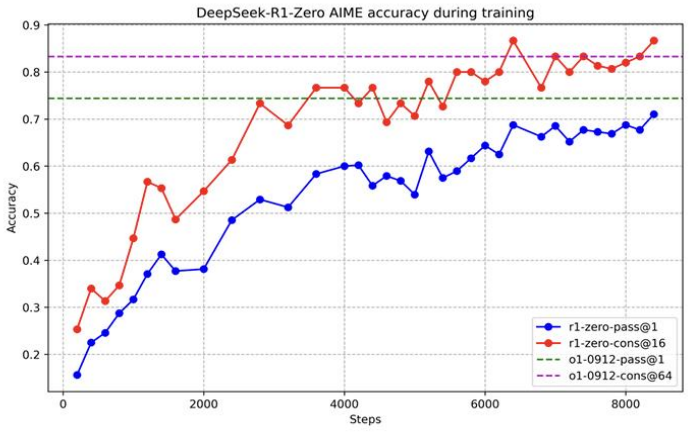
GRPO 通过去价值模型依赖与动态组评分两大革新，将传统 PPO 算法的单步优势计算转化为批量相对评估，使训练效率提升。同时激活硬件层面的新型需求：**需支持动态计算图（Dynamic Computation Graph）以加速可变长度思考链（128-4096 Token 弹性伸缩）、稀疏激活内存管理（95%未激活路径仅保留元数据）以及批内并行比较单元（如英伟达 H100 的 Transformer Engine**

指令集优化）。

图表 7：DeepSeek R1 Zero 推理能力显著提升，达到了与 OpenAI-o1-0912 相当的性能水平

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
OpenAI-o1-0912	74.4	83.3	94.8	77.3	63.4	1843
DeepSeek-R1-Zero	71.0	86.7	95.9	73.3	50.0	1444

Table 2 | Comparison of DeepSeek-R1-Zero and OpenAI o1 models on reasoning-related benchmarks.



数据来源：DeepSeek R1，金元证券

pass@1 为模型一次性通过测试，衡量模型准确度；cons@64 为模型多次测试，为模型答案一致性考量

三、效率提升≠需求下降

DeepSeek 在算法层面的突破显著降低了训练阶段的算力门槛——根据官方披露，R1 模型仅使用 2048 块 NVIDIA H800 GPU（算力成本约 558 万美元）即完成 14.8 万亿 Token 训练，较同类千亿参数模型的典型配置（通常需 5000+GPU）减少 60%硬件投入。这一效率提升主要源于动态稀疏计算架构（单 Token 激活 5.5%参数）与低

秩压缩技术（KV 缓存减少 90%）的协同作用。

然而，算法效率的提升正在加速 AI 应用的规模化落地，进而催生总算力需求，"降本→普及→增量"的螺旋上升效应爆发：

- Post-training 微调算力激增： Post-training 阶段的海量微调（如企业日均执行数万次任务）会持续消耗可观算力，高效微调技术（如 LoRA）虽将单任务能耗压至预训练的 1%-5%，但规模化部署下的总量仍对算力基础设施提出高并发、低延迟需求。
- 云端推理并发量指数增长：AIGC 应用推动云端推理 QPS（每秒查询量）持续攀升，用户要求响应延迟<100ms，驱动高带宽存储与低延迟互连成为刚需。
- 边缘侧长尾需求爆发：通过模型蒸馏技术，DeepSeek R1 能够很容易部署至本地并进行微调，尽管单设备算力需求不敌云端，但总量需求仍大。

本质上，算法优化并非削弱算力产业价值，而是通过重构需求结构打开更大市场空间——从集中式训练向分布式推理延展，从通用计算向场景专用架构升级，最终形成万亿级算力市场的多级增长引擎。

图表 8：更多用户和场景化采用 AI，形成规模化效应

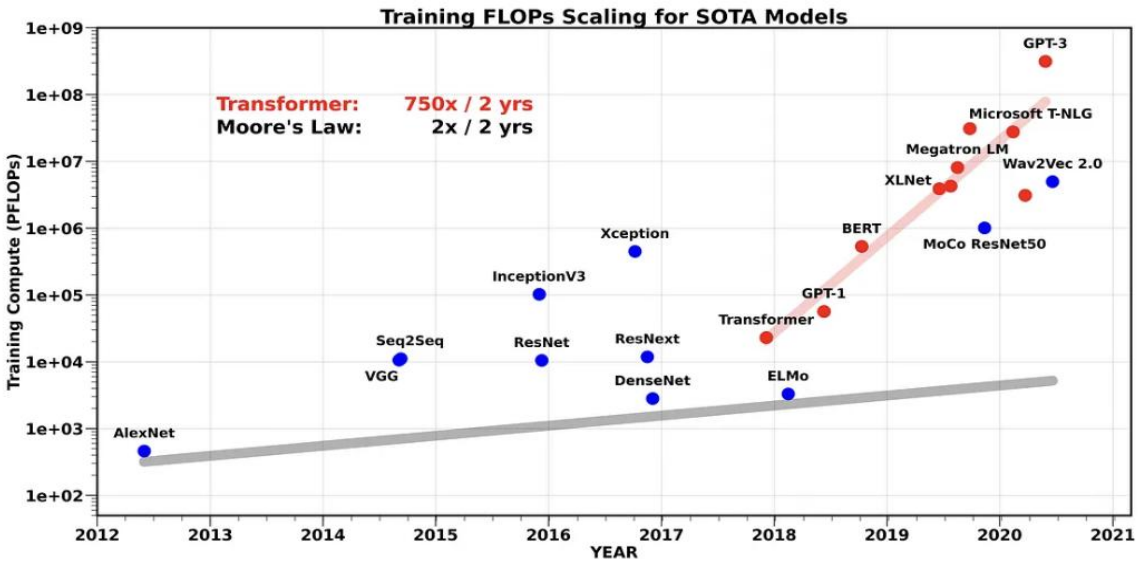
公司/机构	接入时间	合作内容	行业/领域
百度	2025 年 2 月	百度搜索、文心智能体平台全面接入 DeepSeek，优化拍照解题功能，千帆平台上线 DeepSeek-R1/V3 模型	互联网、AI 云服务
腾讯微信	2025 年 2 月 15 日	灰度测试“AI 搜索”功能，接入 DeepSeek-R1 满血版模型，支持联网搜索和隐私保护	社交、搜索
中国电信、移动、联通	2025 年 2 月	三大运营商全面接入 DeepSeek，推动通信产业智能化升级，优化算力成本与代码生成功能	通信、云计算
奇安信	2025 年 2 月	QAX 安全大模型深度接入 DeepSeek，提升威胁研判性能 16%，降低运营成本	网络安全
深圳市龙岗区政府	2025 年 2 月 8 日	政务外网部署 DeepSeek-R1 全尺寸模型，赋能公职人员智能办公	政务
宿迁市政务云	2025 年 2 月	完成 DeepSeek 私有化双区部署，计划构建多参数模型矩阵服务体系	智慧城市、政务
华南理工大学	2025 年 2 月 22 日	本地部署 DeepSeek-R1 满血版，支持复杂推理与私有知识库融合，并向省内院校共享算力服务	教育、科研
钛动科技	2025 年 2 月	全球首批 AI 产业布局企业，应用 DeepSeek 优化算力兼容性	跨境营销、AI 应用
东风汽车	2025 年 2 月	旗下自主品牌车型接入 DeepSeek 全系列大模型，应用于智能交互与车载系统	汽车制造
中国石化、国家能源集团	2025 年 2 月	完成 DeepSeek 私有化部署，推动石油化工、能源领域智能化转型	能源
微博智搜	2025 年 2 月 20 日	内部测试接入 DeepSeek-R1，结合微博讨论内容生成分析结论	社交媒体
清华大学	2025 年 2 月	通过开源项目实现 DeepSeek-R1 满血版本地部署，支持单张消费级显卡运行	教育、技术研发
西城区数据局	2025 年 2 月	推动辖区企业接入 DeepSeek，加速“中国数据街”建设，聚焦数据要素流通与产业生态构建	数字经济、政策规划
奇瑞汽车	2025 年 2 月	星途星纪元 ES 车型接入 DeepSeek-R1，实现语音交互与智能驾驶功能优化	智能汽车
致远互联	2025 年 2 月	协同管理平台接入 DeepSeek，提升企业业务流程自动化与数据分析能力	企业服务

数据来源：公司官网、公众号等，金元证券研究所

四、模型参数量、训练数据持续扩充，高性能算力芯需求仍高

虽然 DeepSeek 在算法层面优化运算效率，但 AI 模型的性能高度依赖训练数据的规模和多样性以及参数量。随着 AI 应用的复杂性增加，训练数据集规模持续扩大。根据 Zhewei Yao、Sehoon Kim 等人研究发现，大型 Transformer 模型中的参数数量呈指数级增长，每两年增加 410 倍，**这也使得采用并行计算、浮点计算能力更强的 GPU、FPGA、ASIC 的需求远胜传统 CPU。**

图表 9： DeepSeek R1 Zero 推理能力显著提升，达到了与 OpenAI-o1-0912 相当的性能水平

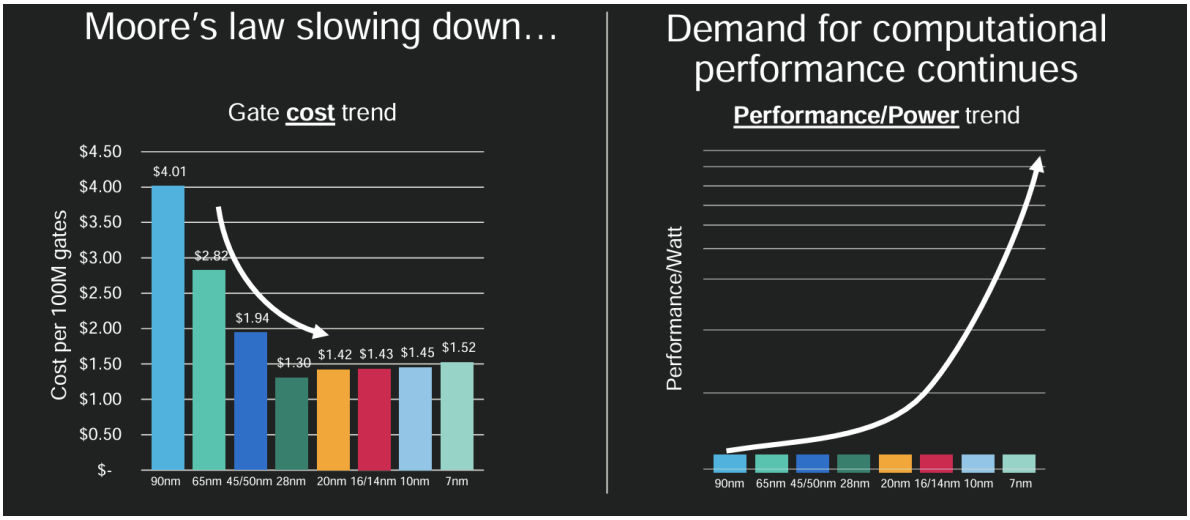


数据来源： AI and Memory Wall，金元证券研究所

传统芯片设计与制造通过缩小 Mosfet 尺寸去提高芯片性能的方式已触达物理极限，量子隧穿效应、短沟道效应以及光刻技术均

在一定程度上使得摩尔定律放缓，高性能芯片的可靠性及功耗成为难题。更重要的是，摩尔定律描述的是芯片的规模化效应，但有不少研究表明随着 FET 的栅长缩小，其成本并未下降。28nm 技术节点后，FinFET 取代平面技术构建了 3D 结构后，成本居高不下。

图表 10： Moore 定律放缓，且规模化效应锐减

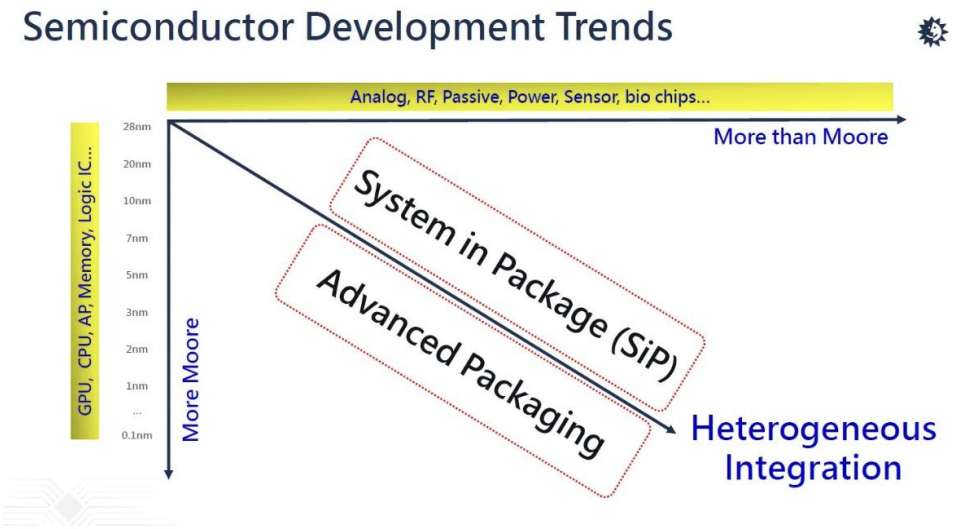


数据来源：Marvell，金元证券研究所

单纯倚仗传统芯片设计与制造通过缩小 Mosfet 尺寸去提高芯片性能的方式效率降低，先进封装技术是后摩尔时代的关键。**随着晶圆技术的进步，片内互连（IC 层级）与系统板级互连（PCB 层级）之间的线宽/线距（Line/Space，L/S）差距日益扩大，导致信号完整性恶化、带宽受限及能效下降。**这一矛盾已成为制约算力芯片性能的关键瓶颈。面临单芯片制程的物理极限及经济效益递减问题，先进封装是“More Than Moore”（超越摩尔）时代的

解决方案。半导体封装不仅提高了系统性能，也是人工智能发展的关键基础。

图表 11：先进封装弥补单芯片制程极限



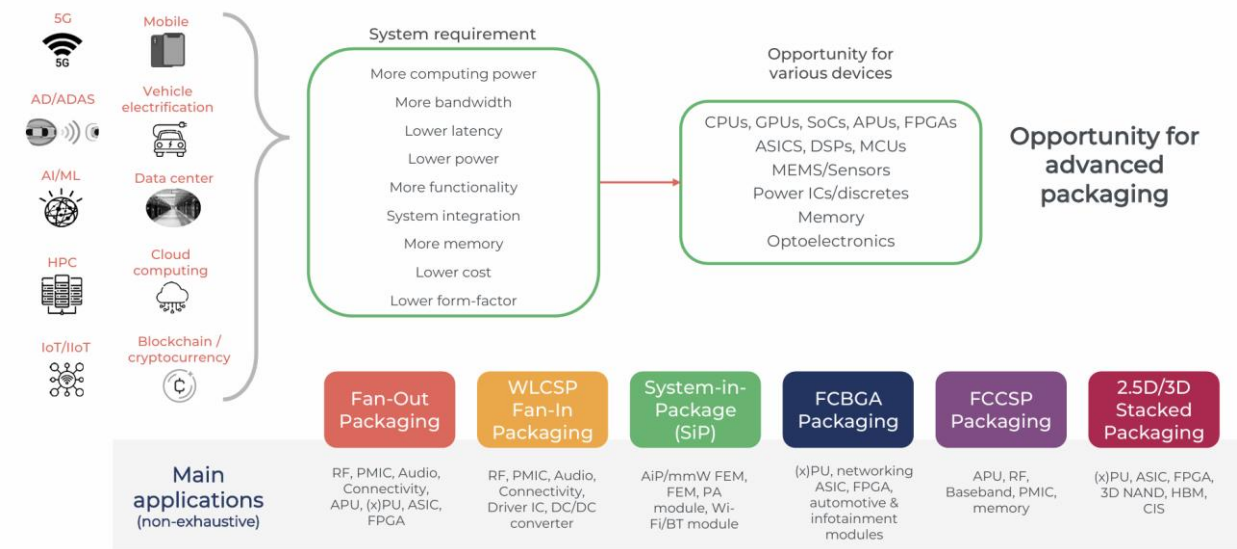
数据来源：日月光，金元证券研究所

系统级线宽/线距瓶颈在 AI 时代更为凸显。计算需求呈现爆发式增长，对 AI 芯片的算力、存储带宽、能效比提出了前所未有的高要求。**然而，系统级线宽/线距瓶颈限制了高速数据在芯片之间、芯片与外部存储器之间高效传输，严重制约了 AI 芯片性能的充分释放。**

与传统的通用处理器（CPU）相比，AI 加速器（如 GPU、ASIC、TPU）在模型需求并行计算以及高浮点运算上具有一定优势。但是，AI 芯片对带宽、延迟、能效等方面要求更为苛刻：

- **极高的互连带宽需求：**AI 计算中，特别是在深度学习、强化学习等算法，通常在 Pre-training（预训练）需要大量训练数据，例如 DeepSeek-V3 基于 14.8 万亿 Token 的数据进行训练。这些数据需要在计算单元之间和存储单元之间频繁交换，对互连带宽提出极高需求。
- **极低的互连延迟需求：**随着 AI 应用落地至不同领域，对于实时性和响应速度在不同应用场景至关重要。例如在自动驾驶、实时语音识别、智能机器人等。AI 芯片系统中，数据需要在计算单元、存储单元等组件之间快速传递，**任何互连环节的延迟都会累积，最终影响系统的整体响应。**
- **极高的能效比需求：**AI 芯片的功耗问题日益突出。随着 AI 模型规模的不断扩大和计算复杂度的不断提升，AI 芯片的功耗也急剧增加，甚至达到数百瓦甚至更高。高功耗不仅增加了散热设计的难度和成本，也限制了 AI 芯片在移动设备、边缘计算等功耗敏感场景的应用。**系统级互连的功率效率是影响 AI 芯片整体能效比的重要因素。系统级互连的电阻损耗、信号传输损耗等都会转化为热量，增加系统功耗，降低能效比**

图表 12：高端 AI 加速器对先进封装技术的需求

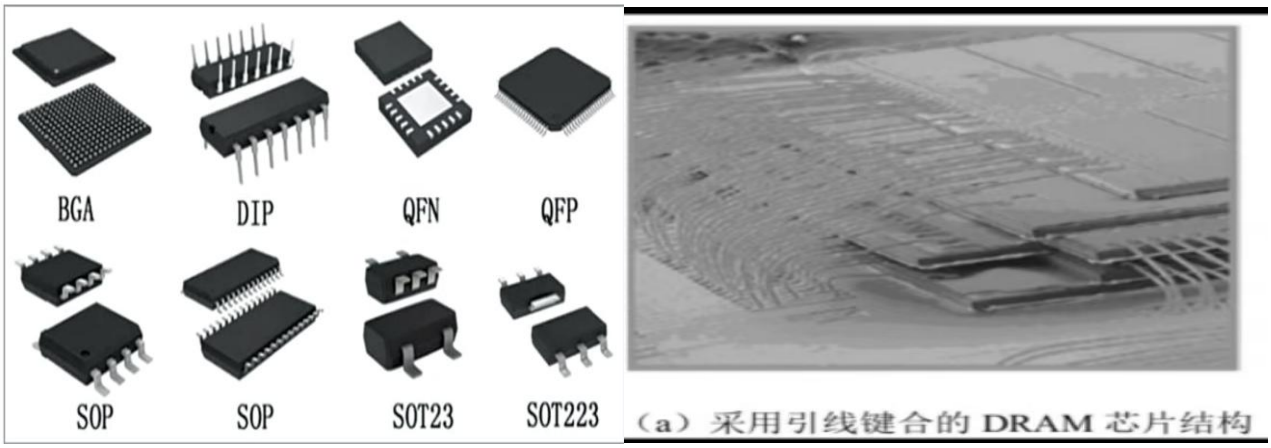


数据来源：Yole，金元证券研究所

五、先进封装实现高性能算力芯片的性能释放

传统封装的工艺相对简单，材料成本较低。且在封装尺寸和互连密度上有限，例如引线键合（Wire Bonding）仅能将芯片的金属导线连接到封装基板中，引线长度较长，寄生电容较大，影响高速信号传输。并且，引线键合的布线密布不足问题在纳米级 IC 尤为凸显。诸如塑料封装、陶瓷封装的热效应较差，导热性及散热降低芯片的，例如 DIP、QFN、QFP 等。

图表 13： 传统封装技术（左为塑料封装及常见封装技术，右图为引线键合）

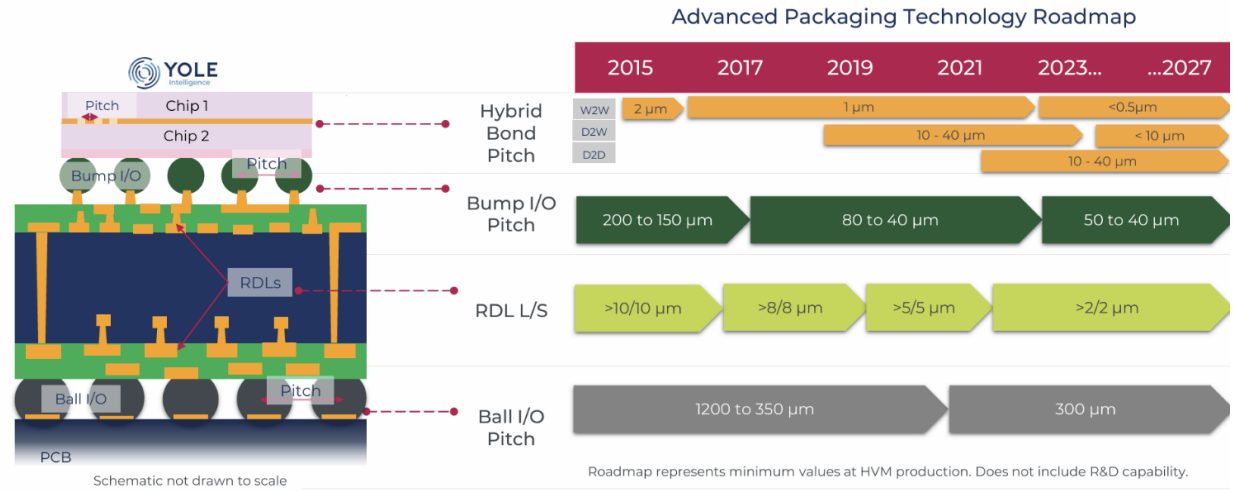


数据来源：electronicsforu、《集成电路系统级封装》，金元证券研究所

先进封装区别于传统的引线键合、塑料封装等封装形式的新型封装技术，**其目标是实现更高密度的互连、更小的封装尺寸、更优异的电气性能和热性能，以及更高的可靠性。**

传统半导体封装技术依赖于印刷电路板（PCB）来进行芯片的连接与集成。PCB 通常作为芯片之间的连接平台，负责将多个芯片互联并提供外部电路连接。然而，随着集成度的提高，PCB 技术面临着尺寸、性能、功耗和成本等方面的限制。并且，PCB 板上走线宽度与凸点及微凸点差异较大，导致 IO 数量及密度限制芯片间或系统互连能力。随着 AI 高性能计算等应用的需求不断增长，传统封装方式无法满足对高带宽、低延迟和小型化的要求。

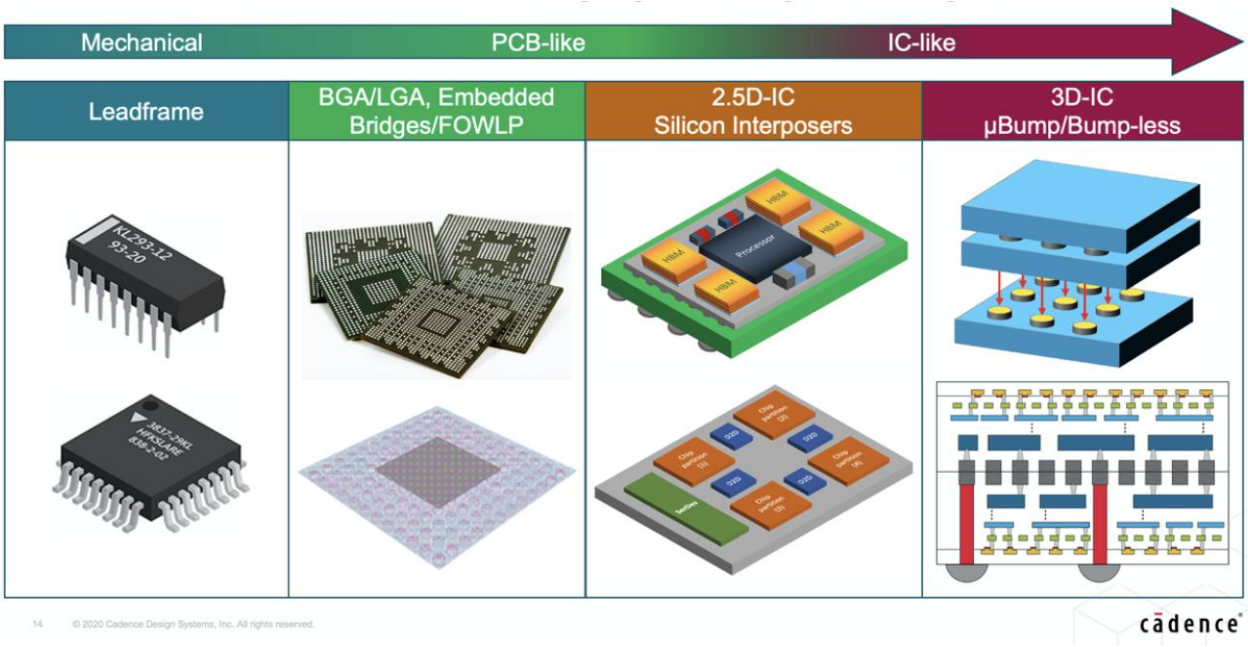
图表 14：IO Pitch 差异



数据来源：YOLE，金元证券研究所

因此，封装技术正逐步从 PCB 的层面，向芯片内部（即 IC 层面）转变。**采用 2.5D 和 3D 封装技术，不再依赖传统的 PCB 作为主连接平台，而是直接将多个 IC 芯片通过转接板（interposer，如硅转接板、玻璃转接板等）进行集成。**这样做不仅提高了芯片之间的带宽，也缩短了信号传输的距离，减少了信号损耗。

图表 15:封装技术演进



数据来源：Cadence，金元证券研究所

引线框架封装（leadframe Packaging）：引线框架封装的核心引线框架是一种金属冲压或蚀刻而成的框架，用于芯片与外部电路的电气连接和机械支撑。芯片通过 WB 技术将芯片上的焊盘与引线框架的内引脚相连，实现信号传输。最终通过塑料或陶瓷等材料进行塑封或气密封装。引线框架封装的优势在于成本低，技术成熟可靠。**但是，I/O 密度瓶颈无法满足高性能算力芯片对高带宽互连的需求。其次，焊线带来的寄生参数限制信号传输速度，延迟较高，且散热性能较差。**

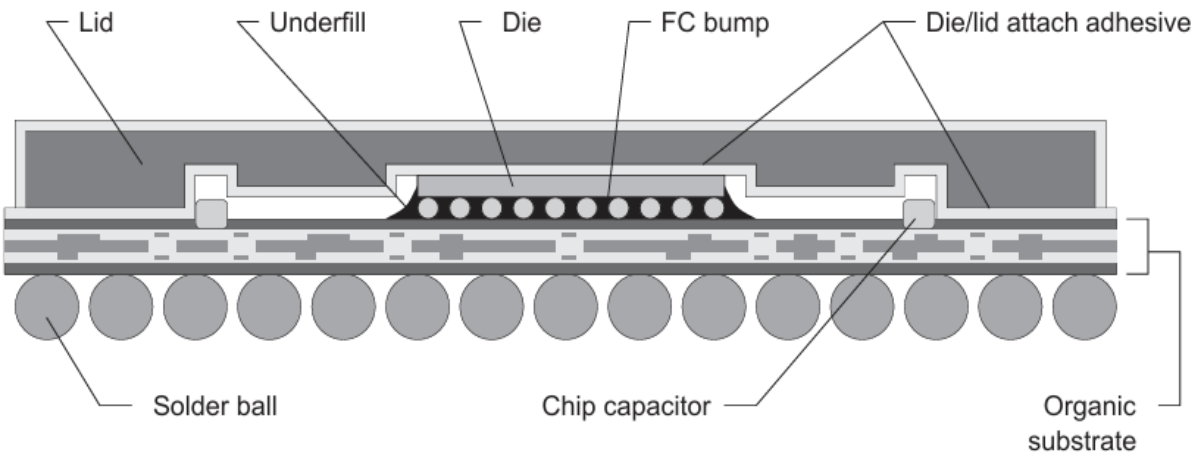
球栅阵列封装（Ball Grid Array，BGA）是对引线框架的改进。BGA 封装取消了传统的引线框架，转而采用再封装基板底部制作

阵列状焊球(Solder Balls)的方式来实现芯片与 PCB 板的互连。芯片与封装基板之间仍然可以通过焊线或倒装芯片（Flip Chip）等技术进行连接。**但是 BGA 的焊球间距较大(通常为 0.8mm,1.0mm,1.27mm），限制了互连密度。随着 I/O 数量增加，封装尺寸也会随之增大。**

FCBGA（Flip-Chip Ball Grid Array）是对 BGA 封装的改进和升级。FCBGA 采用芯片倒装技术(Flip-Chip)将芯片正面朝下键合到基板上，通过微凸点（Micro-Bumps）或焊料凸点(Solder Bumps)实现芯片与基板之间的电气连接。FCBGA 封装的主要结构特点包括:

- 基板（Substrate）:通常采用多层陶瓷或高性能有机材料，比如 ABF（Ajinomoto Build-up Film）材料，提供更高的布线密度和更好的电气性能。
- FC:芯片正面朝下通过微凸点或焊料凸点键合到基板上。
- 焊球:与 BGA 相似,排列在基板底部,用于连接封装体和 PCB。
- 底部填充胶（Underfill）：填充在芯片和基板之间，增强互连可靠性，缓解热膨胀系数不匹配导致的应力问题。

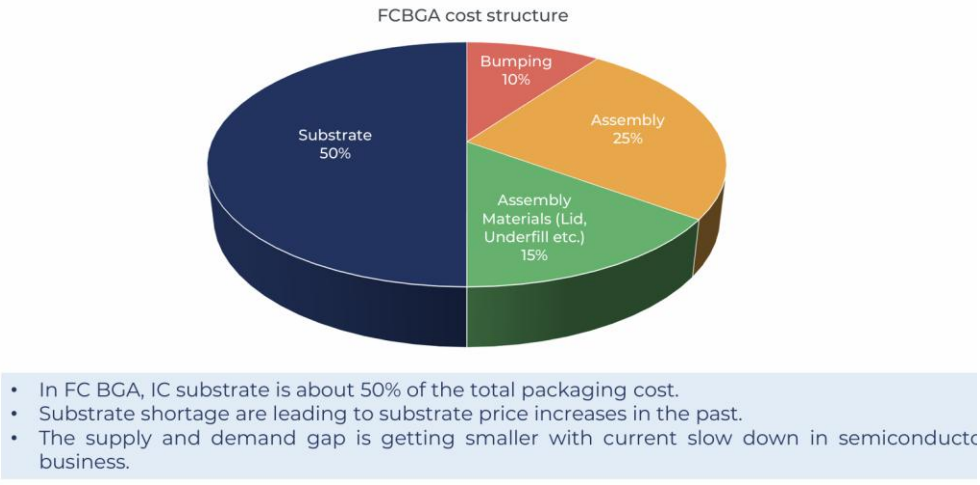
图表 16:FCBGA 结构



数据来源：TI，金元证券研究所

FCBGA 基板通常采用高性能、高可靠性材料，实现芯片与外部电路之间的信号传输。并且在热管理方面，基板材料需要帮助芯片散热，避免过热导致性能下降。根**据 YOLE 数据，封装基板成本约占总成本的 50%。**

图表 17： 基板成本约占 FCBGA 封装成本的 50%

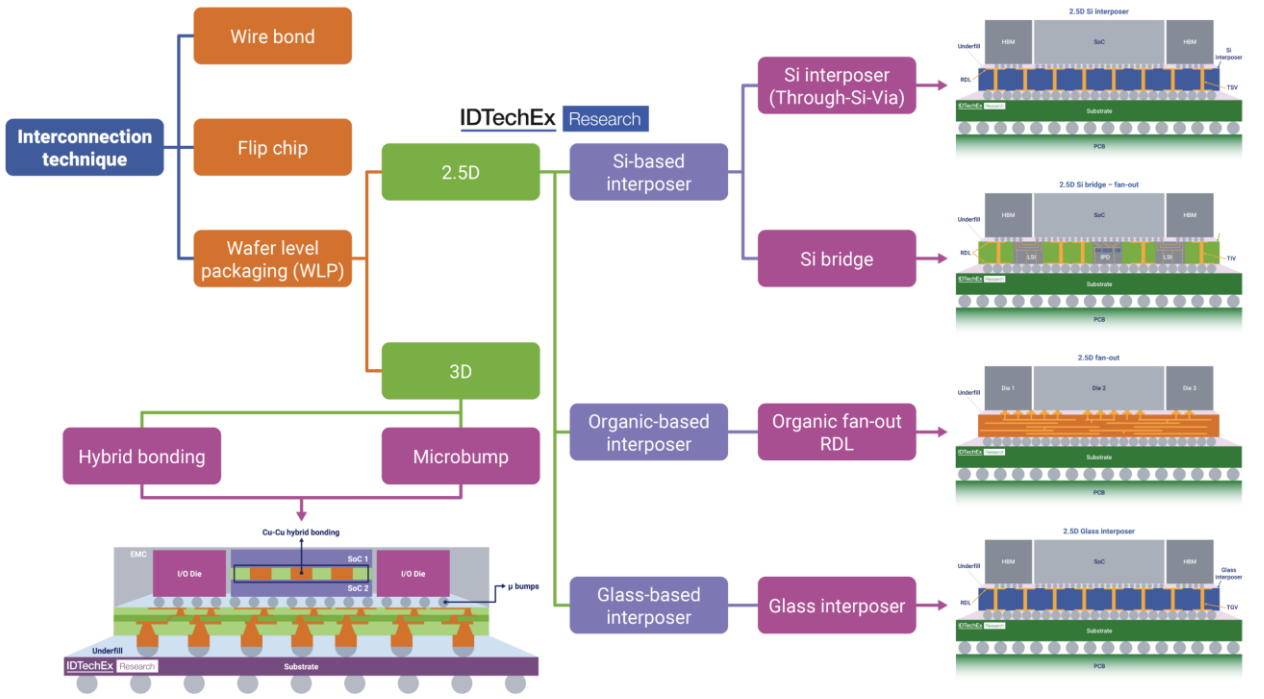


数据来源：YOLE，金元证券研究所

倒装芯片技术消除了 Wire-Bond BGA 中的连接线，采用了更小的微凸点，有效减小了芯片与基板之间的互连间距，提高了互连密度。且芯片与基板之间的信号传输路径更短。**但是，FCBGA 本质仍属于 BGA 一类，其基板布线密度和焊球间距仍然是限制 I/O 数量和带宽的核心原因。**

为了克服 BGA 和 FCBGA 在互连带宽方面的瓶颈，2.5D 封装技术应运而生。**2.5D 封装技术的核心要素在于中介层（interposer）、RDL（Redistribution layer）、硅通孔（TSV，Through Silicon Via）、凸块（Bump）。**

图表 18：2.5D 封装全景

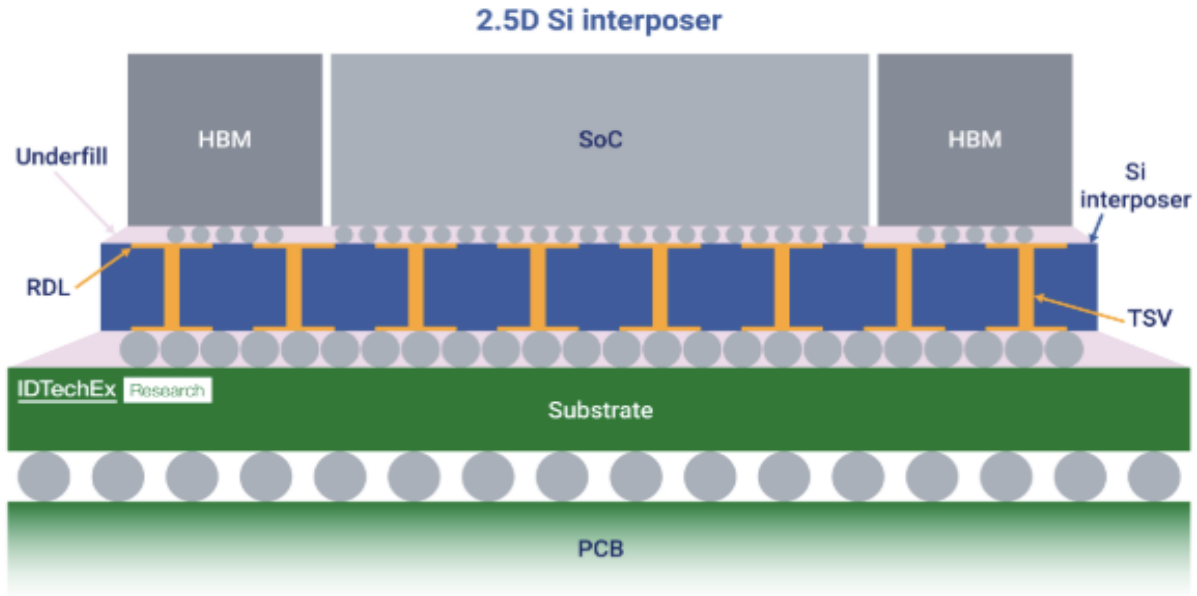


数据来源：IDTechEx Research，金元证券研究所

中介层设置在电子元器件和基板之间作为中介桥梁，实现芯片与元器件维细布线与封装基板稀疏布线之间引脚间距的转换。转接板根据材质划分为无机转接板和有机转接板。无机中介层以 Si、或玻璃、陶瓷为材料，除核心通孔外，还包括底材上下表面的重布线层，利用 TSV 连接上下表面的重布线层。有机转接板采用电镀铜柱连接绝缘材料间的金属线路层。**有机材料的优点在于其介电常数低于硅，有助于降低封装体中的 RC 延迟。**

由于硅中介层能够提供三维布线，芯片之间的连接可以更加紧密，这有助于减少封装的总尺寸。因此，在相同面积下，硅中介层可以实现更多的信号连接，从而提高布线密度。这种紧凑的设计方式对于高性能计算、高带宽应用尤为重要。

图表 19：Interposer 及 2.5D 封装结构



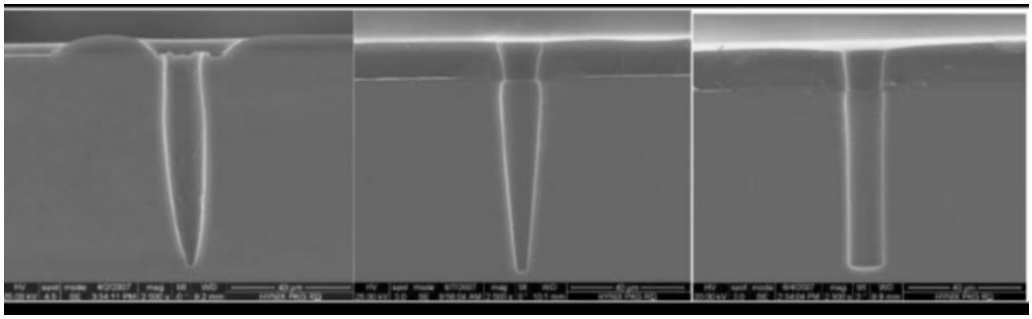
数据来源：IDTechEX Research，金元证券研究所

TSV（Through Silicon Via，硅通孔）是 2D 晶圆平面封装到 2.5D 封装，最终形成 3D 晶圆堆叠封装的关键技术，包括台积电的 CoWoS-R 和新一代高带宽存储器(High-Bandwidth Memory,HBM)均通过 TSV 进行垂直方向上的互连通信。**TSV 工艺难点在于 TSV 刻蚀，填充及背面露头，这也导致了 TSV 的高成本和地良率，降低成本和提高良率成为了 TSV 大规模应用的市场驱动力之一。**

TSV 刻蚀主要方法分为两种，一种是深反应离子刻蚀（Deep Reactive Ion Etching，DRIE），另一种是激光钻孔。激光钻孔除了具有成本优势外，其他方面都难以与深反应离子刻蚀相比。

图表 20：TSV DRIE 刻蚀与激光钻孔的性能对比（左下 1 为激光钻孔、2 为 RIE、3 为 DRIE 工艺）

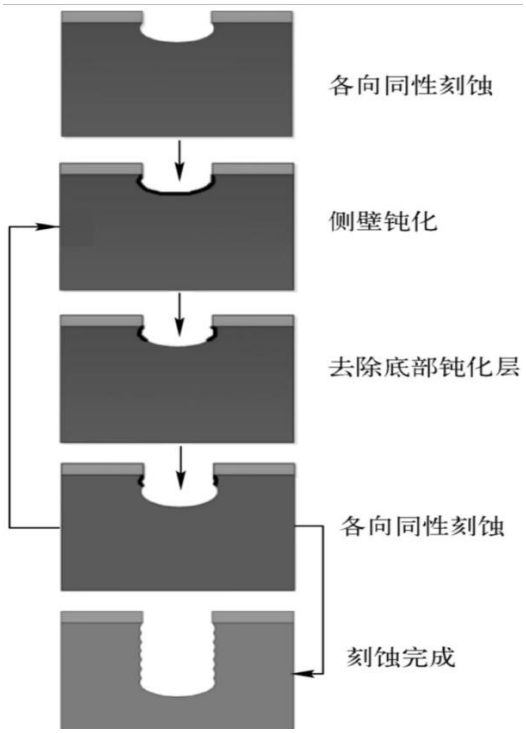
制作工艺	孔径	深宽比	粗糙度	均匀性	通孔角度	生产效率	成本
深反应离子刻蚀	<5μm	20	好	好	-90	高	高
激光钻孔	>20μm	>20	差	差	85	低	低



数据来源：《三维集成电路制造技术》王文武，金元证券研究所

深反应离子刻蚀形成硅通孔主要采用 Bosch 工艺，单个刻蚀周期分为刻蚀和钝化两个步骤，先通过SiF_6刻蚀气体各向同性刻蚀至一定深度后，为了制造高深宽比的通孔需要在 TSV 内通过C_4F_8钝化气体形成一层保护层，然后重复多个刻蚀周期，最终形成高深宽比 TSV 通孔。

图表 21：DRIE 刻蚀步骤



数据来源：《三维集成电路制造技术》王文武，金元证券研究所

刻蚀后需对 TSV 通孔沉积绝缘层对硅衬底进行电气隔离，形成绝缘层材料一般采用SiO_2、SiN_x和其他聚合物。SiO_2、SiN_x等无机介质材料一般使用 PECVD（等离子增强化学气相沉积）工艺沉积，沉积速率较高，且工艺温度低、台阶覆盖性能好。

为了实现信号导通，TSV 填充一般使用电阻率较低的铜。铜不仅有较好的导电性，而且熔点高，后续工艺温度匹配性较好。但是铜的扩散率较高。为了抑制铜扩散需要一层扩散阻挡层阻挡铜原子与硅或氧化硅接触，业界一般采用物理气相沉积（PVD）的 Ti 层。

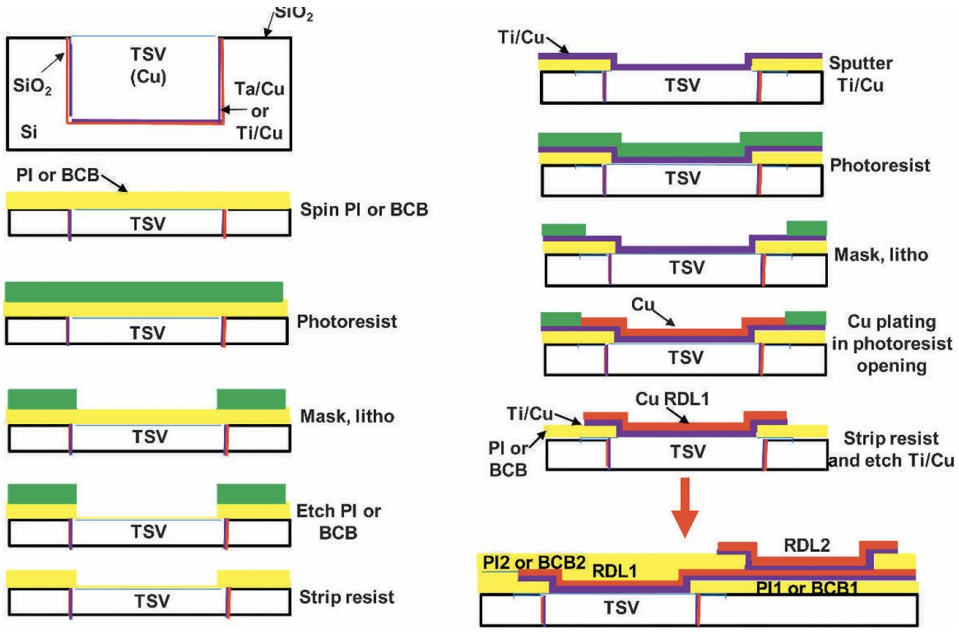
填充方面，TSV 镀铜主要有均匀镀铜及自下向上镀铜方式。均匀镀铜广泛应用于低成本的晶圆级封装，通过合理调配镀液抑制剂和添加剂，在电镀是通孔侧壁和底部均匀生长，但均匀镀铜并不适用于小孔径、高深宽比的 TSV 填充。自下向上的镀铜工艺可以满足无电镀孔洞的镀铜需求。通过使用特殊电镀添加剂、电镀设备及经过特殊设计的电场，在电镀时可减慢通孔外表面铜的沉积速率，加速通孔内部铜的沉积。最后使用化学腐蚀或化学机械抛光 CMP（Chemical Mechanical Polish）工艺来去除芯片表面的铜覆盖层和扩散阻挡层。

RDL 重布线层通常采用铜作为导线材料，介质材料可以使用二氧化硅(SiO_2)、氮化硅（SiN）或其他聚合物，其主要用于在中介层上进行精细布线，RDL 可以将芯片上的 I/O 焊盘重新分布到中介层更大的区域，方便与封装基板或 PCB 板的连接。并且，RDL 可

以构建低阻抗的电源分配网络，为芯片提供稳定的输入/输出电压。

RDL 一方面实现多芯片间的互连，另一方面实现芯片 I/O 与 TSV 位置的再匹配。RDL 的工艺较为复杂，涉及较多关键制造设备。**RDL 的制造工艺包括种子层沉积、RDL 光刻、RDL 电镀、去胶、种子层刻蚀、PI 钝化等步骤，重复上述步骤可以形成多层 RDL 互连。**精度越高的光刻和刻蚀技术，线宽/线距越小.先进的 RDL 工艺可以实现微米级，甚至亚微米级的线宽和线距。

图表 22：多层 RDL 工艺流程

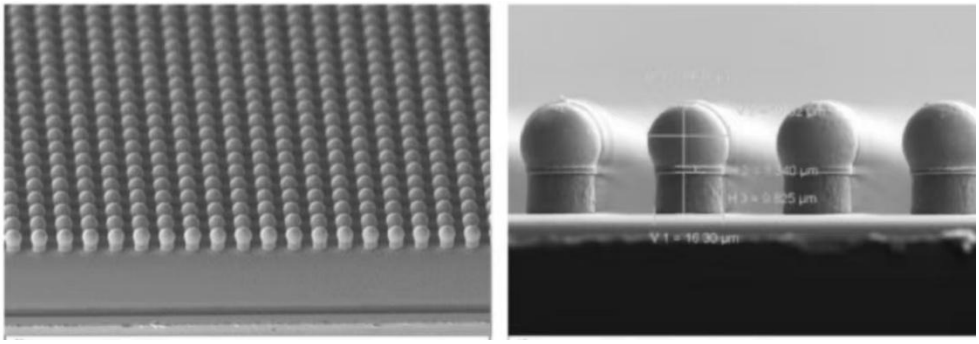


数据来源：《Redistribution layers (RDLs) for 2.5D/3D IC integration》，金元证券研究所

凸块（Bump）/微凸点（Micro-Bumps）主要用于芯片与中介层以

及中介层和封装基板的连接。微凸点通常采用铜柱凸点（Copper pillar bump）结构，有时会在铜柱顶部覆盖一层薄的焊锡层。微凸块尺寸更小，间距更密。凸块的制造与 RDL 类似，同样需要包括光刻、金属沉积形成金属种子层、电镀和回流焊等步骤

图表 23：微凸点



数据来源：《三维集成电路制造技术》王文武，金元证券研究所

实际上，各大厂商在 2.5D 封装技术上组合中介层、RDL、硅通孔、凸块技术，以满足不同客户需求。

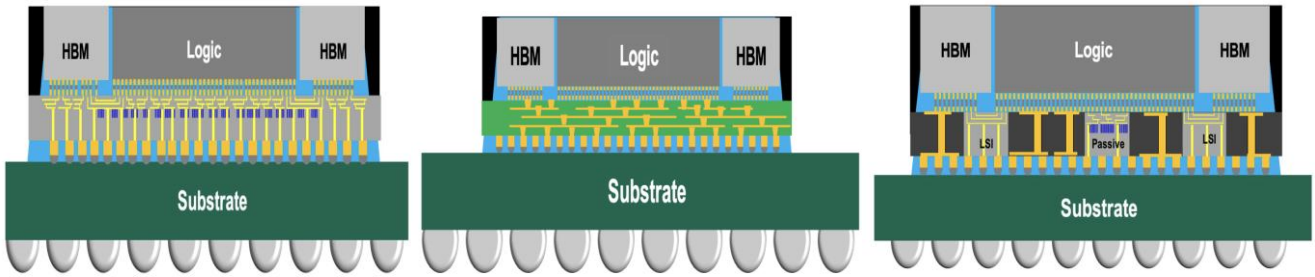
以台积电的 CoWoS-R、CoWoS-L、CoWoS-R 为例：

- CoWoS-S 是台积电最为正统的封装技术，它结合了 TSV、RDL、Interposer、微凸块/C4 凸块。TSV 集成在硅中介层中，用于垂直连接中介层顶部的芯片与底部封装基板。RDL 位于硅中介层顶部，负责水平布线，将芯片 I/O 重新分布到 TSV 位置。中介层采用**被动硅中介层**（无晶体管），提供高密度互连和热稳定性。微凸块（μBumps）：用于芯片与中介层之间的连接（间

距 40-50μm）。C4 凸块：用于中介层与基板之间的连接（间距 100-200μm）。整体性能及成本均较高。

- CoWoS-R 主要利用 RDL interposer 层完成互连，无需硅中介层，多层高密度 RDL 替代硅中介层，直接布线在有机基板上。
- CoWoS-L 更像是 CoWoS-R 与 CoWoS-S 的中间体。通过 LSI（Local Silicon Interconnect）实现高布线密度 Die-to-Die 互连。LSI 可以在多个产品中具有多种连接结构，比如 SoC-SoC、SoC-Chiplet、SoC-HBM。中介层采用**局部硅中介层（硅桥）**与**有机基板**的混合结构。**其优势在于仅在需要高密度互连的区域使用硅（如 HBM 与 GPU 的连接），其他区域用低成本有机材料。且在逻辑芯片的下方可以迈入额外被动元器件的能力。**

图表 24: CoWoS-S\CoWoS-R\CoWos-L



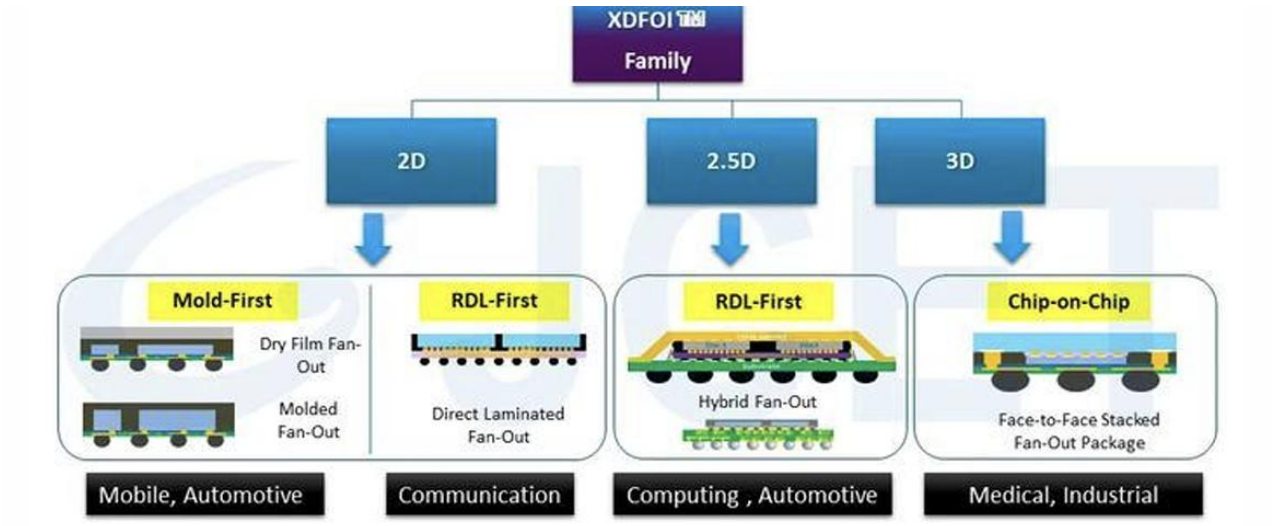
数据来源： ***TSMC****，金元证券研究所*

国内 OSAT 厂商长电科技在 2021 年面向 Chiplet 异构集成应用推出以 2.5D TSV-less(无 TSV)的新一代封装技术，XDFOI(多维扇

出封装集成)。在设计上，长电科技主要通过多层 RDL 来实现高密度走线，其线宽/线距最小可达 2μm。

并且**根据 JCET 公开信息，XDFOI 使用先 RDL，后 Chip 的方式。**RDL-first 工艺最大的优势在于对良率的显著提升。对于传统的 Chip-first 工艺来说，由于已知合格芯片（KGD）已经嵌入，后续加工过程中的各种缺陷都将导致芯片失效，特别是在金属层较为复杂且层数较多的情况下，最终良率往往不能满足要求。而 RDL-first 工艺可以首先完成线路排布，通过测试手段选择合适的区域进行芯片倒装，从而实现大幅度提高产品良率的目的

图表 25:CoWoS-S\CoWoS-R\CoWos-L

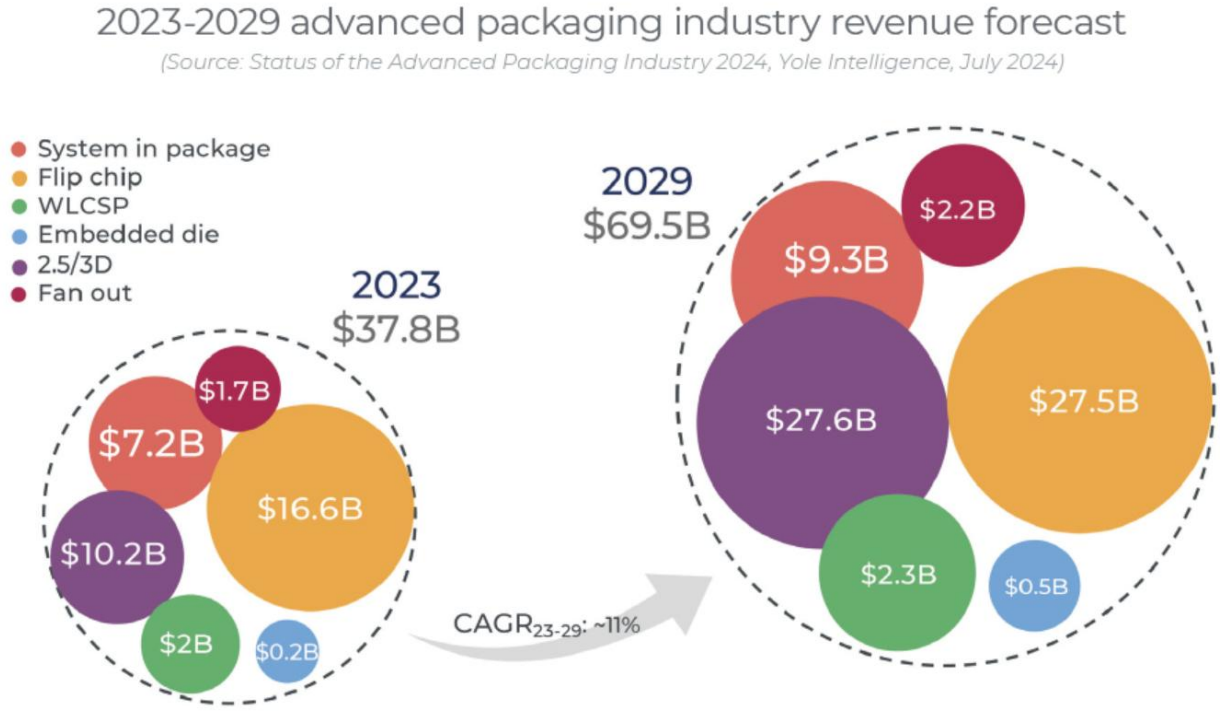


数据来源： JCET，金元证券研究所

DeepSeek V3/R1 的发布及其开源及部署的便利性，很有可能推动 AI 在不同场景化的加速落地。虽然 DeepSeek 通过优化算法减

少冗余计算，提高模型的训练及推理的计算效率，但随着 AI 渗透率及垂直领域发展很有可能带来高性能算力芯片的需求增长。而当前 IC 内部线宽/线距与系统级线宽/线距的巨大差异限制了算力芯片的带宽及表现。先进封装技术，尤其是 2.5D 层面可满足不同场景下高带宽、低延迟的需求。2.5D 的关键技术包括硅中介层、TSV、RDL layer、Bumps，各大厂商通过组合这些技术达到客户的不同场景化需求。根据 YOLE 预测，2023 年全球先进封装营收约 378 亿美元，占半导体封装市场的 44%；2024 年增长至 425 亿美元，至 2029 年，先进封装营收有望增长至 695 亿美元，年复合增长率 11%，其中 2.5D/3D 封装渗透率最快。

图表 26：先进封装市场 2023-2029 年期间，有望达到 11%的复合增长，且 2.5D/3D 渗透率提升



数据来源： YOLE，金元证券研究所

六、投资建议

我们认为，2.5D 封装涉及多项前道工艺设备，包括气相沉积设备制作 RDL 层及 TSV 扩散阻挡层；刻蚀设备也应用于高深宽比的 TSV 刻蚀应用。先进前道设备必不可少。

另外，封装基板材料对封装性能、热管理、电气特性和可靠性构成较大影响，尤其是高性能芯片需要更强热管理及机械支撑。ABF 材料具有较低的介电常数和低损耗特征，能够在高速、高频的应用中有效传输信号，减少延迟和反射。此外，ABF 材料适用于多层堆叠和复杂的信号线路设计。

先进封装 OSAT 直接受益，随着 HPC、AI 芯片及云端等垂直领域应用落地，对于高带宽、低延迟的封装技术要求更高。受益于 2.5D/3D 的高速增长及渗透率，有望迎来强增长。

图表 27：相关公司

公司分类		公司代码	公司简称	摘要
前道设备厂商	气相沉积设备	002371.SZ	北方华创	公司为客户定制的 UBM/RDL 金属沉积设备、TSV 金属沉积设备、TSV 刻蚀设备及工艺已经实现了在国内主流先进封装企业的批量生产，并不断获得客户的重复采购订单
		688072.SH	拓荆科技	公司在现有PECVD设备基础上，针对先进封装领域晶圆的特殊性， 采用独特的加热盘、传片平台等设计，**开发了反应温度在80℃-200℃范围内（通常反应温度在260℃-550℃范围内）的低温薄膜沉积设备，可以沉积低温的SiN、TEOS 等介质薄膜材料，并在先进封装领域实现量产应用。**
	电镀设备	688082.SH	盛美上海	公司基于自主知识产权的前道铜互连电镀设备Ultra ECP map及 电镀工艺，将该技术进一步延伸到先进封装湿法设备领域，**成功开发了先进封装电镀设备、三维TSV电镀设备和高速电镀设备，填补国内空白并形成批量销售。** 同时布局多款后道先进封装工艺设备，技术优势明显。
	刻蚀设备	688012.SH	中微公司	公司ICP技术设备产品类中的8英寸和 12 英寸深硅刻蚀设备Primo TSV200E®、Primo TSV300E®在晶圆级先进封装、**2.5 维封装和微机电系统芯片生产线等成熟刻蚀市场继续获得重复订单的同时，在300mm 的3D芯片的硅通孔刻蚀工艺上得到成功验证，并在欧洲客户300mm微机电系统芯片的生产线上获得认证机台的机会。**
OSAT	封测厂商	600584.SH	长电科技	公司集合长期各项先进封装技术，推出面向Chiplet的高密度多维异构集成平台（XDFOI），利用协同设计理念实现了芯片成品集成与测试一体化机，涵盖2D、2.5D、3D Chiplet集成技术。2023年1月公司宣布其XDFOI工艺已经按计划进入稳定量产阶段，同步实现国际客户4nm节点多芯片系统集成封装产品出货。
		002156.SZ	通富微电	公司在2021年搭建2.5D/3D封装平台VISionS及超大尺寸FCBGA研发平台。公司超大尺寸2D+封装技术、3D堆叠技术、大尺寸多芯片Chip Last封装技术已经验证通过。
材料厂商	基板材料	002436.SZ	兴森科技	2022 年，公司进一步 拓展到ABF载板领域，完善先进封装载板产业布局。公司 FCBGA 封装基板项目已完成珠海和广州基地一期产能建设，数字化管理系统建设持续推进，产品良率接近海外龙头，已完成部分国内标杆客户的工厂审核和产品认证，成为内资工厂中为数不多具备 FCBGA 封装基板业务量产能力的厂商之一。

数据来源：公司官网，金元证券研究所

风险提示

1）2.5D\3D 封装及其他先进封装难度较大，良率有待改善，或影响利润

2）前期设备投入及研发成本较高

3）AI 应用落地速度不及预期

金元证券行业投资评级标准：

增持：行业股票指数在未来 6 个月内超越大盘；

中性：行业股票指数在未来 6 个月内基本与大盘持平；

减持：行业股票指数在未来 6 个月内明显弱于大盘。

金元证券股票投资评级标准：

买入：股票价格在未来 6 个月内超越大盘 15%以上；

增持：股票价格在未来 6 个月内相对大盘变动幅度为 5%～15%；

中性：股票价格在未来 6 个月内相对大盘变动幅度为-5%～+5%；

减持：股票价格在未来 6 个月内相对大盘变动幅度为-5%～-15%