“韬定律”新论文解读 20260906

AI 驱动半导体周期复苏与“套定律”先进封装技术演进研报

摘要

- “套定律”通过逻辑折叠与混合键合压缩信号延迟，麒麟 2026 已实现 1.5μm 键合间距，2027 年目标 1μm，未来三年有望达到 720nm。
- 逻辑折叠显著优化功耗：麒麟 2026 晶体管密度提升 55%，但通过缩短线长、多核降压及热感知布局，NPU/GPU/CPU 功耗分别降低 66%/58%/41%。
- 散热需求随 AI 性能提升而增加，主动散热成趋势：微泵方案散热效率比被动方案高 3 倍以上，厚度减薄 70%，艾为电子、飞荣达等产业链公司受益。
- AI 驱动半导体周期全面复苏：2026 年效应从 GPU 扩散至模拟芯片，A 股模拟龙头 AI 业务占比从 5%升至 30%，科创芯片板块表现强劲。
- AI 服务器重心转向互联系统：GPU 价值占比从 95%降至 60%，存储占比升至 30%，CoWoS 先进封装成为解决“内存墙”的核心瓶颈。
- 先进封装工艺壁垒极高：TSV 需高深宽比刻蚀，RDL 需双大马士革工艺，Hybrid Bonding 可实现 5μm 以下 Pitch，互联密度较 MicroBump 提升量级。
- 晶圆代工厂面临价值重估：AI 拉动稼动率满载及先进制程需求，中芯国际、华虹半导体等公司 PB 估值远低于全球同行，具备补涨空间。

Q&A

华为最新发布的关于“套定律”（τ-law）的论文，其核心观点是什么？该定律与传统的摩尔定律有何区别？

该论文的核心在于阐述“套定律”（τ-law）的可行性，并回应业界对其关键技术——逻辑折叠（Logic Folding）所引发的散热问题的疑虑。 与摩尔定律通过不断微缩晶体管的特征尺寸（空间维度）来提升芯片性能不同，“套定律”在晶体管尺寸微缩受限的背景下，转向缩减信号通过芯片关键路径所需的特征延迟时间（时间维度），即压缩 τ 值。实现这一目标的核心技术是逻辑折叠，它通过混合键合（Hybrid Bonding）技术，将原本平铺的电路折叠成微小的垂直互联结构，

从而有效缩短信号的传输距离。

论文中提及的麒麟 2026 和下一代麒麟 2027 芯片，在混合键合技术上有哪些具体的指标和发展规划？

麒麟 2026 芯片已实现 1.5 微米的混合键合间距，能够达成 5,000 万个垂直互联。下一代产品麒麟 2027 的目标是将键合间距缩短至 1 微米左右。根据技术路线图，预计未来三年内，混合键合的间距有机会追平顶层金属层的间距，即达到 720 纳米，从而实现芯片内金属层间距与混合键合间距更优的比例关系。

逻辑折叠技术如何解决业界普遍担忧的散热问题？论文中提出了哪些具体原因来解释功耗的降低？

论文指出，尽管逻辑折叠会增加晶体管密度，但在同等计算需求下，通过优化设计反而降低了功耗。以麒麟 2026 为例，其晶体管密度相较前代产品从每平方毫米 1.55 亿个提升至 2.38 亿个，增幅达 55%。然而，在性能对齐的真实负载测试中，其 NPU 功耗降低了 66%，GPU 功耗降低了 58%，CPU 功耗降低了 41%。 功耗降低主要源于以下三个方面： 第一，通过逻辑折叠节省数据“搬运”功耗。现代 SoC 约 90%的功耗为动态功耗，其中数据搬运（在不同模块间传输比特）的能耗远超计算本身。逻辑折叠将原本需要跨越数百微米的传输路径，通过亚微米间距的混合键合缩短至几微米。评测显示，折叠路径上的线长平均缩短约 20%，关键路径上最多可缩短 70%。这减少了导线电容，从而显著降低了通信密集型模块（如 NPU、GPU）的功耗。 第二，利用多核并行运作实现降压。逻辑折叠释放的芯片面积可容纳更多晶体管，从而实现更广泛的并行处理。工作被分摊到更多的核心上，使得每个部分能以更低的电压和更慢的时钟频率运行。由于动态功耗与电压的平方成正比，降压带来了显著的功耗节省。例如，麒麟 2026 的 NPU б л а г о д а р я折叠带来的晶体管富余，核心数增至四个大核，使其能在 0.7 伏电压下提供 70 TOPS 的算力，而上一代产品在提供不到一半算力时就需要 0.85 伏的电压。 第三，通过热感知布局主动规避热点。在设计阶段，会有意识地避免折叠高功耗电路，并在各层之间采用交错布置的方式，防止局部热点在空间上彼此叠加。这种选择性地在收益最大的关键路径上应用折叠，并结合热感知布局，而非盲目折叠所有电路，从结构上提升了热管理水平。

在具体的芯片模块上，逻辑折叠技术带来的功耗和性能改善效果如何？

逻辑折叠技术对通信密集型模块的改善效果尤为显著，模块越需要数据来回搬运，其带来的回报就越大。 以 NPU 为例，作为最热的模块之一，在相同的 29 TOPS 计算需求下，麒麟 2026 的时钟频率可以下降 63%，工作电压从 0.85 伏降低到 0.55 伏，最终实现了 66%的功耗降低和 73%的功率密度下降。 GPU 的表现紧随其后，在与麒麟 9,030 Pro 相同的 61 FPS 运行条件下，其电压下降了 200 毫伏，功耗随之降低了 58%。这些功耗的梯度下降与通勤路径的缩短直接相关。

从产业链角度看，“套定律”（τ-law）的实践和推广将对哪些环节产生深远影响？

“套定律”的实践将对半导体产业链的多个环节带来显著影响。首先，该技术路线在客观上会消耗更多的晶圆，因此中芯国际、华虹等晶圆代工厂将深度受益。其次，逻辑折叠技术的核心是混合键合，随着键合间距的不断缩小，对先进封装技术和设备的需求将大幅增加。因此，提供混合键合设备的厂商（如拓荆科技）以及具备先进封装能力的封测厂将扮演至关重要的角色，并从中受益。

根据最新的“韬定律”论文，通过热感知布局等方式优化功耗是否意味着散热需求会减少？热管理在其中扮演何种角色？

功耗优化并不意味着散热需求的减少，热管理依然是核心约束。这主要基于两点：第一，论文中实现的功耗下降是单位算力的功耗下降，而非整机功耗的下降。垂直堆叠架构通过将长距离的横向布线改为极短的垂直路径，显著缩短了信号传输距离，从而优化了单位算力任务的功耗。然而，这种优化出的功耗空间会被用于增强端侧 AI 能力、运行更复杂的 AI 模型等计算任务，导致整机功耗随 AI 性能提升而增加，整体散热量也随之增加。第二，该论文反而凸显了局部散热问题的重要性。堆叠架构的热流密度显著高于平面架构，当多层高功耗计算单元在空间上重叠时，单位面积产生的热量会急剧增加。这些热量需要穿过多层硅、键合界面等材料才能传导出去，容易在中间形成局部热点和热量阻塞。虽然堆叠两层时或许可以通过设计层面缓解，但未来当堆叠层数增加到三层、四层甚至更多有源层时，通过布局规避热点重合的难度将越来越大，局部散热问题会愈发关键。因此，热管理不会成为“韬定律”发展的阻碍，但随着终端 AI 能力的增强和堆叠层数的增加，整机功耗和局部热流密度都将上升，散热问题依然是发展的核心。

在整机功耗增加和局部热流密度攀升的趋势下，散热技术将如何演进？主动散热方案相较于被动方案有何优势？

在整机功耗与局部热流密度攀升的背景下，主动散热将成为新的趋势。传统平面 SOC 的平均热流密度约为每平方厘米 15 瓦，而堆叠架构下的热点热流密度可能飙升至每平方厘米数百瓦。在这种情况下，由传统石墨、铜管、VC 均热板等组成的被动散热方案难以满足需求，引入主动散热方案成为必然。主动散热方案的散热效率比被动方案高出 3 倍甚至更多，并具备多重优势：首先，它能动态调节散热速度，通过微泵单元结合温控算法调节泵速；其次，它具有小型化、轻薄化和耐弯折的特点，例如微泵厚度仅为 0.8 至 1.5 毫米，比 VC 均热板薄 70%以上，更适用于手机、平板等曲面或紧凑型设备；此外，还具备噪音低、功耗低等优点。因此，随着手机、平板、笔记本电脑等产品的终端 AI 能力增强，主动散热方案将在这些场景中得到更广泛的应用。



目前主动散热技术的主要路径有哪些？产业链中有哪些值得关注的公司？

当前主动散热技术主要围绕微泵工艺展开，具体包括液泵和气泵两种路径，两者目前都处于商业化的初期试验和研发阶段。液泵的技术原理是通过高压驱动芯片驱动压电微泵，使冷却液在膜片内循环流动以吸收热量。气泵则是一种微风冷工艺，通过高频振动直接对芯片吹风以降温。在产业链层面，建议关注艾为电子和飞荣达。艾为电子是液泵驱动芯片的国内独家供应商。整个液泵模组由驱动芯片、压电微泵和液冷膜片三部分构成，而驱动芯片的全球供应商仅有两家，分别是海外的 Boreas 和国内的艾为电子，其市场卡位优势显著。同时，艾为电子在风冷工艺上也有布局。飞荣达则专注于整个微泵模组的制造。随着主动散热技术的演进和市场渗透，其远期市场空间广阔。考虑到 9 月华为可能发布搭载“超晶率”的终端新品，艾为电子和飞荣达都有望受益于这一主题催化。

从宏观角度看，当前半导体周期呈现出哪些特点？AI 在其中扮演了怎样的角色？

AI 是驱动本轮半导体周期上行的核心动力，并重构了行业格局。历史上，半导体行业存在明显的周期性，每轮周期上下行时间通常为一年半到两年。若以全球半导体月度销售额同比增速数据为基准，本轮上行周期始于 2023 年 5 月，至今已持续三年，但各细分领域的复苏节奏并不同步。在前期，即 2023 年至 2024 年，周期的拉动主要由 GPU 及其配套的 HBM、DDR 等存储产品驱动。ChatGPT 问世后，GPU 需求爆发式增长，以英伟达 2026 年第一季度单季收入达 800 多亿美元为例，整个全球半导体市场规模已从过去的 4,000 亿至 6,000 亿美元区间，向万亿美元级别迈进。然而，在 2025 年下半年之前，除 GPU 和存储外，半导体其他领域并未实现真正的上行。直到 2026 年，AI 的拉动效应才全面显现。

AI 对半导体行业的拉动效应是如何从核心芯片扩散到更广泛领域的？以模拟芯片为例，其景气度变化情况如何？

AI 的拉动效应在 2026 年开始全面扩散至半导体行业的更广泛领域，模拟芯片是其中的典型代表。由于模拟芯片的下游应用非常分散，涵盖汽车、工业、消费电子、能源等，各占约 20%多的份额，因此在 2023 年至 2025 年 AI 需求主要集中于算力芯片时，模拟芯片行业并未出现高增长。然而，进入 2026 年，AI 开始拉动模拟芯片的多个下游领域复苏，例如工业、能源以及计算类应用，从而带动了模拟芯片整体的景气度回升。从海外龙头公司 TI、ADI、MPS 等的业绩说明会来看，它们对整个周期的展望已开始释放强劲信号。同时，观察 A 股的模拟芯片龙头公司，AI 相关业务在其收入中的占比也出现了显著提升，从 2025 年的约 5%增长至 2026 年的 30%。这表明，AI 的拉动作用已从最初的少数核心环节，渗透到了原先距离 AI 较远的半导体细分领域。

A 股半导体上市公司在本轮 AI 驱动的周期中表现如何？与消费电子、PCB 等板块相比有何差异？

A 股半导体上市公司在本轮周期中的复苏节奏相对滞后，但自 2026 年以来表现强劲。整体而言，A 股上市公司距离 AI 核心产业链较远，在 2022 年和 2023 年，许多公司仍在处理存货减值和无形资产等问题，经营较为困难。2024 年，部分公司开始出现复苏迹象。到了 2025 年，如果与电子行业内的消费电子、PCB 等板块相比，半导体板块整体的业绩释放强度较为一般。然而，进入 2026 年，情况发生显著变化。AI 最终的瓶颈落在了半导体上，其强劲的拉动作用全面显现。截至目前，科创芯片板块年初至今的涨幅已达到 67%，显著超过了电子行业内其他细分板块，印证了 AI 对半导体板块的强力驱动。

从 2026 年第一季度财报来看，半导体行业呈现出怎样的复苏态势？AI 在其中扮演了什么角色？

从 2026 年第一季度的财报数据来看，半导体行业内多个细分领域，包括材料、设备、制造和设计等，均已出现显著的同比增长，显示出明确的复苏迹象。可以认为，本轮半导体周期是由 AI 技术的发展所拉动的。

根据 Rubin 的 OM 成本构成图，AI 服务器的价值构成发生了哪些变化？这揭示了行业发展的何种趋势？

根据 Rubin 的 OM 成本构成图，AI 服务器中 GPU 的价值占比正在下降。在 2023 年 A100 服务器刚推出时，GPU 的价值量占比高达 90%至 95%；而现在，这一比例已降至约 60%。与此同时，其他组件的价值在提升，例如存储的价值占比已提升至 25%到 30%，而 PCB、光模块等组件的价值也基本实现翻倍增长。这一变化表明，AI 服务器的发展路径已不再仅仅依赖于 GPU 的单点升级，互联系统的升级已成为最核心的驱动因素。

从英伟达 GPU 的升级路径来看，其发展呈现出哪些特点与挑战？这如何印证了互联系统的重要性？

英伟达的 GPU 升级路径显示，其制程从 A100（2020 年发布）的 7 纳米升级至后续 H100、B100 等产品的 4 纳米，并在 4 纳米制程上停留了大约三年（2022 年至 2025 年），计划于 2026 年开始升级至 3 纳米。这一过程反映出 GPU 升级的难度日益增大，成本也随之升高。同时，一个关键认知是 GPU 大部分时间在等待数据而非执行计算。因此，当 AI 服务器发展到当前阶段，单纯依靠 GPU 性能提升已非唯一路径，系统集成与互联效率成为新的关键因素。这解释了为何 AI 的发展能够带动半导体产业的方方面面，从过去单一依赖 GPU 性能，转变为现在需要整个互联系统协同进步的局面。

AI 从训练阶段向推理阶段的演进，对半导体产业链的具体需求带来了哪些变化？为何说半导体是当前产业升级的核心瓶颈？

AI 从训练向推理的演进，对半导体产生了更广泛且多样化的需求，包括存储用量的增加、CPU 的升级、模拟器件的升级，以及光模块普及带来的光芯片需求增长。产业升级的核心瓶颈最终都落在了半导体及其相关产业链上。其原因是，在全球工业化和自动化发展至今的背景下，多数行业的扩产相对容易，唯独半导体行业的扩产难度最大。这主要体现在投资金额巨大、扩产决策极为谨慎，以及产业链条过长（涵盖上游材料设备、中游制造各环节及下游封测等），通常需要一年半到两年的扩产周期。然而，市场需求不会等待产能完全准备好后才出现，这种需求上行与供给不足的错配，往往会引发行业稼动率的提升。当稼动率满载后，若供给依然跟不上，则会出现价格上涨。下游客户为确保供应，可能会锁定产能、抢夺订单，从而引发“overbooking”（超额预订），进一步推高价格和产业链相关公司的业绩。企业在确认真实需求后才会启动新一轮扩产，从而形成了特有的行业周期。

贵司如何看待 AI 重构半导体产业的投资节奏和机会传导路径？

我们认为，AI 对半导体产业的重构是一个层层传导的过程。我们自 2025 年 8 月起开始逐步推荐相关受益产业链。最初的关注点是存储，因为推理应用带来了存储需求的显著上行，我们推荐了兆易创新及存储模组等公司，并持续关注国科微等标的，认为存储是价值量持续提升的长期方向。到 2026 年 2 月，我们开始重点推荐设备领域，主要基于国内两家主要存储厂商庞大的扩产规划，这背后既有 AI 拉动全球存储需求的因素，也有国产化率提升的重要议题，我们看好拓荆科技、中微公司、北方华创等整个设备产业链。进入 3 月份，我们开始强力推荐国产算力相关产业链，因为观察到国内已形成一个从 AI 模型到算力需求的良性循环。以智谱、Kimi 等为代表的国产大模型快速发展，其 token 调用量在全球排名前五中占据四席，这直接导致了国产算力设计公司的订单上修，并进一步传导至产能环节。近期，我们重点推荐晶圆制造厂的价值重估，因为无论是成熟制程稼动率的提升，还是先进制程与先进封装的发展，最终都会利好 Fab 厂，为其带来估值上行的空间。

从全球市场来看，晶圆制造厂的估值水平呈现何种趋势？这对 A 股和港股的相关公司有何启示？

全球范围内，晶圆制造厂正在经历一轮价值重估。以先进制程的代表台积电为例，其市盈率已接近 12 倍。在其他领域，如硅光领域的 Tower Semiconductor，其市净率也达到了约 11 倍。即便是成熟制程厂商，如世界先进，其估值也重估至 3.5 到 4 倍 PB 的水平。相比之下，A 股和港股的相关公司，如中芯国际和华虹半导体，其港股的 PB 估值基本在 0.8 倍到 1 倍左右。鉴于存储行业已经历了业绩与估值的双重重估，我们认为在其他逻辑驱动下，晶圆制造厂板块也应迎来一轮

价值重估。

什么是“韬定律”（Tao's Law）？其技术本质和核心思想是什么？

“韬定律”中的“韬”是一个电路时间常数，它决定了信号的传播速度。这个常数值越小，芯片的运行速度就越快，能效也越高。简而言之，其目标是让电流跑得更快。实现这一目标的核心途径是通过先进封装技术来缩短芯片之间的信号传输时间。“韬定律”的本质在于，当国内在先进制程发展（如从 N+1 到 N+2 再到 N+3）因缺少 EUV 光刻机而面临压力时，我们无法像海外那样单纯依赖光刻机将晶体管刻得更细（从 7nm 到 2nm）。因此，当无法在单一维度上缩小晶体管时，我们选择通过更智能的芯片集成方式，即让信号传输的距离变得更短。这代表着一种发展路径的转变，从单纯依赖制造工艺的进步，转向制造、封装、设计、互联等环节的系统性协同合作。

市场对于“韬定律”的可行性存在哪些担忧？华为的实践经验提供了怎样的例证？

市场对“韬定律”的主要担忧在于它是否能成为像摩尔定律那样经过长期验证、可重复、可预测的普适性工程规律，以及其工程极限（如能否达到 1.4 纳米水平）和整个产业生态的接纳与配合问题。然而，华为的实践已经证明了其可行性。华为已通过该路径实现了 381 款芯片的量产，覆盖通信、计算、传感等多个应用领域。大家熟知的麒麟芯片和升腾芯片，都是通过堆叠、先进封装等技术形态实现了批量生产和性能的代际迭代，并且能够在手机、服务器等多个终端产品上实现大规模出货。这表明，在中国这个占全球半导体需求三分之一的巨大市场中，举国之力推动“先进封装+制程”同步发展的路径是可行的，庞大的市场体量本身就能够带动产业链各环节的协同发展。

“韬定律”强调先进封装的重要性，这是否意味着先进制程的价值会被削弱？

并非如此。虽然先进封装在“韬定律”中至关重要，但这并不意味着先进制程的价值会降低。要将芯片性能推进到例如 1.4 纳米的水平，仍然需要依赖于最先进的制造工艺。尤其在国内，从 N+1 推进到 N+2 乃至 N+3 的过程中，要想真正对标海外的先进水平，必须在现有最先进的工艺基础上发展。我们无法单纯依靠先进封装技术，来完全弥补与海外已发展到 2 纳米制程之间的差距。因此，先进封装与先进制程是相辅相成、协同发展的关系，而非相互替代。

封装技术名词繁多，例如 QFN、BGA 等，其核心发展方向是什么？

众多封装技术名词可以暂时忽略，关键在于理解其核心发展方向。与算力芯片通过微缩制程增加晶体管密度、存储芯片通过微缩制程增加电容密度以提升性能类似，封装技术的核心发展方向是增加单位面积的 I/O 通道数。这一目标旨在匹配算力与存储芯片性能提升后日益增长的通信需求。 封装技术演进的路径清晰地

体现了这一方向。早期采用引脚（Lead Frame）方式，随着芯片性能提升，引脚数量不足成为瓶颈。为增加通道数，引脚被做得更细，并从双边布局发展为四边布局。然而，引线过细易断，且边长有限，进一步提升密度面临极大挑战。为突破此限制，业界转向倒装（Flip-Chip）技术，利用芯片的整个表面积来排布凸点，每个凸点即一个连接通道。后续的发展方向便聚焦于缩小凸点间距和凸点尺寸，从而在单位面积内集成更多的凸点，实现更高的互联密度。

在 AI 时代，为何必须引入 CoWoS 这类先进封装架构？它与传统封装相比有何不同，其核心工艺环节包括哪些？

在 AI 领域，引入 CoWoS（Chip-on-Wafer-on-Substrate）架构是为解决“内存墙”问题，即算力芯片与存储芯片之间通信的延迟和功耗瓶颈。尽管存算一体是终极目标，但由于算力芯片（如 2nm）与存储芯片（如 10nm 以上）在制程、工艺（晶体管 vs. 电容）上存在巨大差异，短期内难以实现。因此，产业选择“近存计算”（Near-Memory Computing）作为过渡方案，其核心思想是让存储与计算单元尽可能靠近，并增加两者间的互联通道数以提升带宽和整体性能。 CoWoS 架构正是为实现“近存计算”而设计的。以 HBM 与 GPU 的集成为例，两者通过各自的 PHY 区域进行通信，这些区域布满了高密度的连接线路。为实现高带宽，这些线路的线宽线距要求达到 0.4 微米级别，远超传统 IC 载板二三十微米的工艺能力。因此，必须引入新的架构。CoWoS 采用硅中介层（Silicon Interposer），利用 65nm 或 55nm 的代工制程来刻蚀 0.4 微米的线宽线距，从而满足 AI 芯片的高密度互联需求。 与传统封装相比，CoWoS 架构的关键差异体现在三大核心工艺环节：

1. 硅通孔：在硅片上垂直钻孔，实现上下芯片间的直接通信，大幅缩短信号路径，降低功耗。HBM 的堆叠也采用了相通的 TSV 技术。
2. 重布线层：在中介层上制作多层精细的金属布线，其工艺难点在于实现极细的线宽线距。
3. 微凸点（Micro-bumping）：在单位面积内制作数以百万计的微小凸点。其技术挑战在于确保所有凸点的高度一致性，并防止相邻凸点之间发生短路。

这些工艺环节大量借鉴了前道工艺，如光刻、曝光和刻蚀，但精度要求介于传统封装与前道制程之间。台积电正是通过在先进封装领域长达十余年的投入（目前研发人员已超 4 万人），构建了从前道制程到后道封装的整体技术壁垒，使得 CoWoS 产能成为分配其先进制程产能的关键。

硅通孔技术在先进封装中的具体作用和制造流程是怎样的？

硅通孔技术的核心作用是在硅片上实现垂直的电气互连，从而打通上下层芯片，使信号能够快速地从上层传输至下层。这种垂直连接方式极大地缩短了传输路径，避免了传统绕线带来的路径过长和功耗增加问题。例如，在 HBM（高带宽内存）的堆叠结构中，TSV 技术是实现多层 DRAM Die 之间高速、低功耗通信的关键。 TSV

的制造流程通常在硅片上进行，尽管业界也在探索在玻璃等其他材料上实现类似工艺。其基本步骤如下：首先，对作为基底的硅片进行清洗；随后，通过刻蚀等工艺在硅片上钻出微小的通孔，实现物理上的贯通。

请详细阐述 TSV（硅通孔）的制造工艺流程及其中的关键技术难点？

TSV 的制造流程始于晶圆清洗，随后通过热氧化或 PECVD 工艺生长一层氧化硅。接着，使用涂胶设备在氧化硅层上均匀涂布光刻胶，再利用光刻机进行曝光，在光刻胶上形成开口。之后，通过刻蚀工艺去除开口处暴露的氧化硅薄层。完成上述步骤后，采用深硅刻蚀设备刻蚀 TSV 的凹槽。此环节的一个核心技术难点在于实现高深宽比（High Aspect Ratio），即在孔径极小的情况下达到较深的刻蚀深度，这对刻蚀设备提出了极高的要求。 值得注意的是，初始刻蚀的 TSV 并不会完全贯穿晶圆。以一片厚度接近 800 微米的硅片为例，台积电的硅中介层厚度可能仅为 100 微米，后续甚至会减薄至 50 微米。因此，TSV 孔洞是在后续工序中，通过晶圆减薄设备从背面进行研磨，直至暴露出 TSV 内部填充的铜层，才能实现贯通。 在 TSV 凹槽刻蚀完成后，使用去胶设备移除剩余的光刻胶。随后，在凹槽内部先生长一层氧化硅作为绝缘层，再沉积一层钛铜（Ti-Cu）或钽铜（Ta-Cu）层。该金属层在半导体工艺中通常用作铜的阻挡层或隔离层，以防止铜原子向周围的硅中扩散。接下来，通过电镀设备在凹槽内填充铜。填充完毕后，利用 CMP（化学机械抛光）设备将晶圆表面多余的铜去除，避免后续制程中可能发生的短路。至此，单个 TSV 的结构基本完成。在 HBM 等产品中，TSV 的数量非常庞大，其数量、孔径与深度的比例共同决定了加工的整体难度。

RDL（重布线层）的作用是什么，其制造工艺，特别是多层 RDL 的实现方式和技术挑战有哪些？

RDL（Redistribution Layer）的主要作用是在芯片上实现复杂的电学连接。以芯片俯视图为例，白色的圆点代表凸点，而下方的走线即为 RDL。在芯片堆叠键合时，上下两层芯片的凸点需要对齐连接，这只能实现垂直方向的点对点导通。然而，芯片内部复杂的电路设计要求实现任意两点间的连接，例如将第一行第一个触点与第二行最后一个触点连通，这就必须通过 RDL 走线来实现。 RDL 的结构类似于 PCB 上的铜线，但其线宽和线距要精细得多，并且通常是多层结构。每一层 RDL 由垂直连接的通孔和水平的金属走线构成。制造多层 RDL 的工艺极具挑战性，其难点在于保持每一层的平坦度。例如，第一层 RDL（RDL 1）在制作完成后可能并非完全平坦，或许存在右高左低的情况。那么在生长第二层 RDL 时，工艺上就必须进行补偿，实现左高右低，以确保整体的平坦性。这种工艺控制需要长期大量的实验摸索。 RDL 的制造工艺主要有两种：电镀和双大马士革（Dual Damascene）。当线宽线距在 5 微米及以上时，通常可以采用电镀工艺，例如使用 PSPI（光敏聚酰亚胺）等材料。但在先进封装领域，线宽线距普遍在 2/2 微米、1/1 微米甚至更小，此时必须采用双大马士革工艺。 双大马士革工艺流程如下：首先在已完成 TSV 的晶圆上通过 PECVD 生长一层氧化硅，然后涂布光刻胶并进行光刻，接着刻蚀氧化硅，再去除光刻胶。随后，沉积钛铜层作为阻挡层，

再电镀填充铜，形成连接上下层的通孔。完成这一步后，需要重复上述流程来制作水平走线：再次生长氧化硅、涂胶、光刻、刻蚀、去胶，然后沉积钛铜层，再进行铜电镀以形成走线。最后，通过 CMP 工艺将表面多余的铜抛光去除，至此一层完整的 RDL 才算制作完成。若要制作多层 RDL，则需重复整个工艺流程。随着层数的增加，工艺步骤累积，良率控制的难度也随之显著提升。尽管 RDL 工艺与前道工艺在涂胶、显影、刻蚀、电镀、CMP 等方面原理相通，但其在精度等方面的要求有所不同。

芯片封装中凸点工艺的关键性体现在哪些方面，其面临的主要技术挑战是什么？

凸点工艺在先进封装中至关重要，是实现高密度互连的基础。随着封装技术的发展，为了在单位面积内集成更多的 I/O 通道，凸点的尺寸被设计得越来越小，凸点间的距离也随之缩减，这极大地增加了工艺难度。 该工艺面临两大核心技术挑战： 第一，确保所有凸点的高度一致性（等高）。在一颗芯片上会制作大量的凸点，如果其中任何一个凸点的高度偏低，那么在与另一颗芯片进行键合对位时，该凸点将无法与上方的对应凸点有效接触，导致连接失败，最终致使整颗芯片报废。 第二，防止凸点之间的桥接（短路）。由于凸点数量庞大且间距极小，必须在制造过程中精确控制其形态，确保任何两个相邻的凸点都不会发生物理接触或连接，否则会引起电气短路。 在 CoWoS 等高端先进封装流程中，制作凸点是一个非常基础且前置的工艺步骤，其质量直接决定了后续封装的成败和最终产品的可靠性。

如何在 RDL 层上制作凸点，以及凸点技术的发展趋势是什么？

凸点的制作流程始于已完成的 RDL 层。首先进行清洗，然后镀上一层钛层或铜层作为种子层。接着，通过涂胶设备旋涂光刻胶，并使用光刻机进行曝光。曝光后，进行电镀铜，再通过特定设备将焊料涂覆在铜上。随后，去除光刻胶，并刻蚀掉多余的铜层和钛层，以防止连接点之间发生短路。最后进行退火处理，使凸点形成具有一定弧度的“小帽子”形状，便于后续键合过程中的上下对齐，其中键合设备的对准精度是一个关键参数。 凸点技术发展的核心趋势是不断缩小凸点间距，以在单位面积内实现更多的连接通道数，从而提升带宽。技术演进路径从早期的打线封装（Wire Bonding）和尺寸较大的植球，逐步发展到缩小凸点尺寸和间距。具体来看，经历了从尺寸稍大的 C4 Bump，到 Micro Bump，再到最新的混合键合（Hybrid Bonding）。

Micro Bump 与 Hybrid Bonding 在技术指标和应用上有何区别，以及 Hybrid Bonding 的技术原理和工艺流程是怎样的？

产业界通常以 10 微米的 Pitch 作为划分 Micro Bump 和 Hybrid Bonding 的界限。然而在实际应用中，Micro Bump 要做到十几微米的 Pitch 已非常困难，目前台积电及 HBM 产品中的凸点 Pitch 普遍在二十几微米的水平。相比之下，Hybrid

Bonding 技术可以轻松实现 5 微米以下的 Pitch，例如 3 微米，并有潜力向 1 微米甚至更小的尺寸发展。因此，Hybrid Bonding 的连接通道数相比 Micro Bump 能实现一个数量级的提升。这项技术是全球先进封装的发展方向，被应用于多种场景，例如兆易创新、长鑫存储的 Xtacking 技术以及后续的 CDA 方案都采用了这一路径。华为也已在其芯片中采用 Hybrid Bonding 工艺以提升带宽和性能。此外，未来的 HBM 产品也将采用此工艺，但目前因多层堆叠导致的良率下降问题尚未实现量产。不过，对于 1 层、2 层或 4 层的堆叠，该技术已相对成熟，AMD 的部分芯片（SRAM 与算力芯片的连接）、台积电的 3D SRAM 以及英伟达的未来方案均会采用 Hybrid Bonding。 技术原理上，传统 Bumping 是向上生长凸点，通过键合设备对齐后退火，使焊料融化实现互连。而 Hybrid Bonding 则是在芯片表面挖出凹槽，向内填充铜。具体工艺流程包括清洗、涂覆光刻胶、刻蚀、镀钛铜种子层、电镀铜填充凹槽，最后通过 CMP 进行抛光。键合时，将两片芯片的铜面对齐，通过加热使铜膨胀实现直接互连。这一工艺使得 Pitch 能够实现指数级的提升。在先进封装中，键合设备是核心环节，Hybrid Bonding 主要使用晶圆对晶圆（Wafer-to-Wafer）或芯片对晶圆（Die-to-Wafer）的键合设备，而传统凸点则多使用 TCB 等键合设备。

以 CoWoS-S 为例，其完整的工艺流程是怎样的？

CoWoS-S 的工艺流程始于准备工作：硅中介层已完成 TSV、RDL 层和凸点的制作，而上方的 SOC 和 HBM 芯片也已制作好相应的凸点。海力士与台积电的紧密合作便是一个实例，海力士将其 HBM 的逻辑 Die 交由台积电代工，以确保双方凸点的 Pitch 等工艺标准能够精确匹配。 具体的封装流程如下：

1. **芯片贴装**：在 12 英寸的硅中介层晶圆（此时未切割）上，使用 Die-to-Wafer 键合设备将 SOC 和 HBM 芯片精确放置并对齐。
2. **底部填充**：芯片键合后，凸点之间存在空隙，需要进行底部填充工艺，将填充材料注入并固化。
3. **临时键合与减薄**：将整个结构固定在一个临时载板（Temporary Carrier）上。由于 TSV 初始并未完全打通，需要使用减薄设备对硅中介层的背面进行研磨，直至 TSV 暴露出来。
4. **背面处理**：在中介层背面制作 RDL 层和凸点。
5. **解键合与切割**：完成背面处理后，将临时载板移除。随后，对晶圆进行切割，将其分割成独立的“SOC+HBM”芯片单元。
6. **最终封装**：将切割好的芯片单元翻转，贴装到 ABF 载板上，并再次进行底部填充。接着，放置散热片（Heat Sink），并使用塑封料（Molding Compound）将整体包裹起来，形成最终的 CoWoS 产品。

整个流程不仅工艺步骤繁多，对精度要求极高，并且大量使用了与前道工艺类似的设备，只是在精度要求上略有不同。

在中国大陆市场，先进封装产业的格局、发展驱动力以及相关的投资机会是怎样的？

在全球范围内，由于硅中介层的制造需要前道工艺，AI 芯片的先进封装形成了由晶圆代工厂主导、封测厂为辅的产业格局。然而，中国大陆的情况有所不同。虽然中介层仍在代工厂制造，但除此以外的其他封装环节目前均由封测厂完成。这使得国内封测厂在 AI 产业链中的地位天然高于全球同行。 国内发展先进封装的另一个核心驱动力在于，由于 EUV 光刻机无法引进，前道先进制程的演进依赖于浸没式 DUV 进行多重曝光，难度较大。因此，通过先进封装技术，如增加 I/O 密度和带宽，或将算力芯片与存储芯片进行 3D 堆叠，缩短互联距离，成为提升芯片整体性能的关键路径。这种“芯粒”思路能够弥补前道制程的不足。例如，通过先进封装与前道工艺的结合，预计到 2031 年，国内芯片性能有望达到全球 1.x 纳米的水平，凸显了先进封装在高端应用领域的重要性。 基于以上逻辑，投资机会主要集中在以下几个环节：

1. **封测厂**：这是需要重点关注的板块。核心标的包括盛合晶微，以及长电科技、通富微电、甬矽电子、汇成股份、华天科技等公司。此外，精测电子旗下的武汉精鸿，其 Hybrid Bonding 工艺在全球范围内具备领先性，使其不仅是设备供应商，也成为国内头部的先进封装厂。
2. **设备**：键合设备是先进封装中的核心环节，价值量占比最高。上市公司代表为拓荆科技，此外新益昌也在布局。其他关键设备还包括：盛美上海的电镀设备，华海清科的抛光设备，以及北方华创等前道设备公司也在 TSV 等先进封装设备上进行了大量布局。
3. **晶圆厂**：先进封装需要先进的晶圆支持，且目前国产 AI 芯片的瓶颈在于 Fab 产能。与即将上市、估值可能达万亿级别的长鑫存储相比，中芯国际和华虹公司的股价近期表现滞涨，市值差异巨大，因此具备较高的投资价值。
4. **材料**：先进封装涉及的材料种类繁多，也存在相应的投资机会。