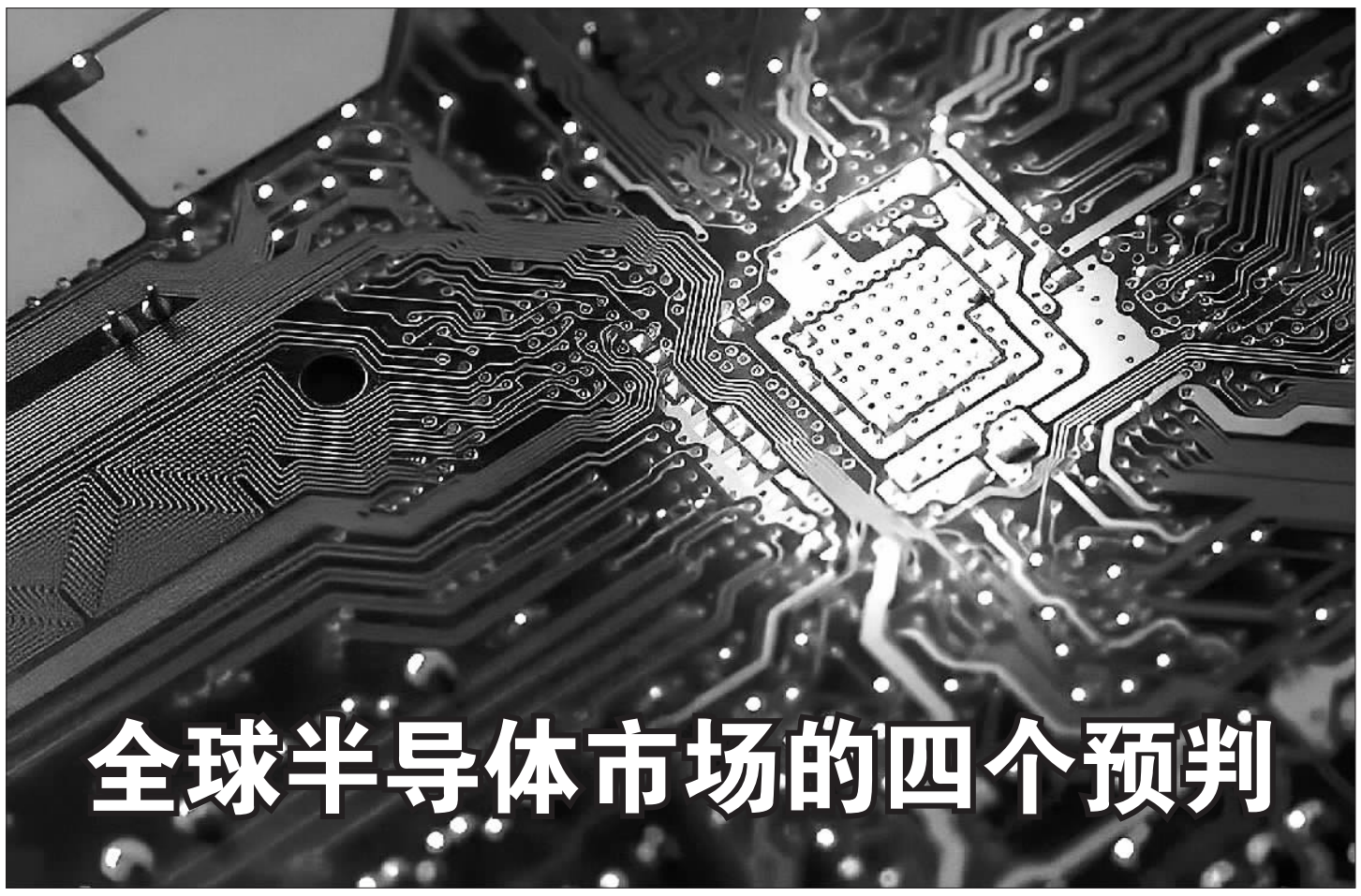

全球半导体市场的四个预判

本报记者 沈丛 姬晓婷

全球半导体市场持久的下行调整期能否画上句号？全球半导体产业又将迎来哪些变化？多家机构和企业给出了自己的答案。

预判一：市场将迎周期性回暖

日前，多家行业协会和市场分析机构做出2024年全球半导体市场回暖的积极判断。

2月6日，美国半导体行业协会（SIA）宣布，2023年第四季度全球半导体产业销售额为1460亿美元，同比增长11.6%，环比增长8.4%。SIA总裁兼首席执行官John Neuffer表示：“2023年年初，全球半导体市场低迷，但在下半年出现强劲反弹，预计2024年市场可实现两位数增长。”SIA预测称，2024年全球半导体产业销售额将增长13.1%。

世界半导体贸易统计协会（WSTS）统计，全球半导体产业在2023年第四季度同比增长6%。市场研究机构Market.us对未来10年的半导体行业进行了展望，总体来看较为积极乐观。该机构认为，全球半导体市场规模有望实现大幅增长，回暖的2024年只是个“开胃菜”，全年的市场总规模将达到6731亿美元。

预计从2023年到2032年，全球销售额将以8.8%的年复合增长率增长，到2032年，预计全球半导体市场规模达到13077亿美元。

1月18日，台积电首席执行官魏哲家在2023年第四季度台积电财报电话会上表示，预测2024年全年除内存外的整体半导体市场将同比增长10%以上，代工行业的增长率预计为20%。

业内专家莫大康在接受《中国电子报》记者采访时表示，2024年半导体市场将迎来周期性回暖，但是回暖程度目前还是未知。对于未来的市场展望，既不应过于乐观，也不应过于悲观。相反，人们应该将当前的挑战视为一种动力，积极应对并推动行业的发展。

预判二：先进制程“节节高”

智能手机和数据中心等高性能集群，是先进制程两大应用场景。

日前，市场调查机构Counterpoint数据显示，2023年第四季度，智能手机出货量同比增长3%，达到3.12亿部，出现复苏态势。Counterpoint预计，2024年全球智能手机出货量有望同比增长3%。

根据台积电发布的2023年第四季度财报，高性能计算（HPC）贡献的收入环比增长17%，占第四季度总收入的43%。黄仁昭表示，高性能计算平台将是2024年增长的最大推动力。

基于该市场需求，3nm及更先进制程成为主要晶圆代工企业的重要竞争领域。

魏哲家表示，几乎所有的智能手机和高性能计算企业都在与台积电合作开发3nm技术。其3nm已成功进入量产阶段，并在2023年下半年实现强劲增长，其销量比上半年多得多。台积电首席财务官黄仁昭在2023年第四季度财报说明会上表示，部分3nm产能可由5nm工具提供支持，这意味着台积电3nm的产能供应能力将有所提升。2024年，台积电3nm的收入贡献也将高于2023年。

与此同时，台积电称，其2nm技术开发进展顺利，器件性能和良品率或超过计划。N2将于2025年实现量产。

三星电子在2023年第四财季报告中指出，其3nm和2nm工艺的GAA架构开发进展较为顺利，近期也收到了一份“2nm的人工智能加速器项目的订单”。

英特尔在2023年赢得了4个18A代工客户，且其代工业务在2024年至2025年期间准备了50多个测试芯片，其中75%将采用英特尔18A。

预判三：人工智能带动GPU持续激增

ADI中国区销售副总裁赵传禹在接受《中国电子报》记者采访时表示，人工智能正以极快的步伐向边缘端发展，得益于自动驾驶和自动化工厂等新应用的出现，人们的目光转向了高效率、实时决策和更安全可靠的运行，进而带动数据处理和智能从云端移向边缘端，这一趋势已蔓延各个行业。智能边缘具备减少延迟、降低带宽需求、提高数据安全性等优势，利用无处不在的检测和人工智能驱动型边缘计算，实时拉近计算、数据存储与数据源之间的距离，数据得以转化为洞察、理解和行动，这能够帮助客户加速实现数字化转型，并对人类和环境保护产生积极影响。

2024年被视为AI PC元年，全球AI PC整机出货量将达到约1300万台。Dell'Oro研究报告预计，2024年GPU产值将继续同比激增70%。

预判四：封测产业2024年“乍暖还寒”

封装市场在2023年遭遇了低谷，但这一现象将在2024年有所改善。长电科技CEO郑力对《中国电子报》记者表示，封测产业2024年“乍暖还寒”，2025年或2026年将迎来较明显的市场上升期。

郑力认为，封测市场的回暖主要是由三个因素带动。首先是消费电子的回暖以及存储市场的助推。郑力表示，随着数据中心的扩张和云计算的普及，存储市场呈现大幅增长，将成为封测产业复苏的推手。

其次是后摩尔时代的不断趋近。郑力认为，随着芯片制程接近物理极限，未来芯片产业需要封测技术扛起性能提升的大旗。这在助推封测市场需求的同时，也对封装技术的多样化提出要求。

最后，新兴市场的涌现也为封测市场注入了新的活力。郑力认为，新能源汽车、光伏发电等绿色能源产业的发展，给宽禁带半导体功率器件的应用落地提供了良好契机。要保证宽禁带半导体器件性能的稳定可靠，离不开先进的封装技术。郑力预测，在2024年，长电科技在宽禁带半导体领域的营收将成倍增长。

近日，英特尔宣布其首个3D封装技术Foveros已实现大规模量产。与此同时，三星也在积极开发其3D封装技术X-Cube，并表示将在2024年量产。

3D封装的理论已经提出多年，但是由于技术难度比较大，能量产3D封装的企业并不多。如今，业内对台积电CoWos等2.5D封装已经供不应求。随着业内对AI芯片算力的需求不断提升，这一现象将很快蔓延到3D封装领域，3D封装甚至将成为AI芯片的制胜法宝。先进制程三巨头在3D封装市场的排位赛也即将开启。咨询公司Yole Intelligence称，未来，全球先进芯片封装市场规模预计从2022年的443亿美元增长到2027年的660亿美元，3D封装预计占四分之一左右的市场规模。

台积电、英特尔、三星竞逐3D封装市场

本报记者 沈丛

台积电最早布局

2022年，台积电成为业内首家量产3D封装的厂商，并将其命名为SoIC。

台积电指出，SoIC支持CoW（芯片在晶片封装技术）和WoW（多晶圆堆叠封装技术）两种不同的封装模式，而这两种模式的结合能够将不同尺寸、功能的芯片进行有效连接，使得芯片在设计过程中具备更高的灵活性。

据了解，目前采用台积电SoIC技术的芯片月产能约为1900片，预计2024年月产能超过3000片，增幅近60%；2027年的月产能有望拉升到7000片以上，是2023年月产能的3.7倍。

此前，英国AI芯片公司Graphcore发布了世界上首颗采用台积电SoIC 3D封装技术的AI芯片，芯片性能提升了40%，并首次突破7nm工艺极限。该款芯片的出现，也展示了3D封装技术在AI芯片领域的巨大潜力。

如今，苹果、AMD等业内龙头企业都成了台积电3D封装的客户。据悉，AMD将在其最新的MI300芯片中采用台积电的SoIC 3D封装技术。苹果计划将SoIC与热塑碳纤板复合成型技术搭配使用，相关产品目前正小批量试产，预计2025—2026年量产。

英特尔实现规模量产

随着AI芯片对3D封装的需求不断增长，仅台积电一家公司的3D封装产能难以满足庞大的市场需求。刚刚实现3D封装量产的英特尔，或将缓解市场的焦虑。

据悉，在此前的2D以及2.5D封装技术中，英特尔基本上都将其用于生产自家产品。但是在3D封装领域，英特尔开始接受外部订单，与台积电展开竞争。

就在不久前，英特尔宣布其首个3D封装技术Foveros已实现大规模量产。英特尔相关负责人向《中国电子报》记者透露，英特尔于去年年底发布的酷睿Ultra处理器已经采用了Foveros 3D封装技术，而此次宣布量产则意味着英特尔可以为客户大批量生产3D封装产品。

此外，英特尔近期发布的2024年财报明确指出，其先进封装代工客户新增三家。业内猜测，这些客户中可能有英伟达，并预计英特尔最快于2024年第二季度正式加入英伟达先进封装供应链行列，为其提供每月高达5000片的产能。

在先进制程领域，英特尔一度落后三星，但是在3D封装领域英特尔却先三星一步实现量产，并同样将在代工市场分一杯羹。

芯谋研究副总监严波认为，英特尔的3D封装技术之所以能快速发展，部分原因是美国建设本土产业集群带来的助推作用。

“虽然英伟达、AMD等公司的AI芯片仍采用台积电的先进封装技术，但目前美国是全球大型芯片设计公司的聚集地。在美国致力于强化半导体供应链的背景下，英特尔的先进封装技术有望迎来本土政策带来的发展红利。”严波说道。

三星蓄势待发

作为代工三巨头之一，三星正在积极开发其3D封装技术X-Cube，并表示将在2024年量产。同时，其为AI芯片开发的最新3D封装技术SAINT也渐行渐近。

X-Cube是三星在2020年的3D封装技术。该技术是将硅晶圆或芯片物理堆叠，并通过硅通孔（TSV）连接，最大程度上缩短了互连长度，在降低功耗的同时提高了传输速率。

2023年，三星推出了3D封装技术SAINT，主要有三种方案：在垂直堆叠SRAM内存芯片和CPU中采用的SAINT S，在CPU、GPU等处理器和DRAM内存中使用的SAINT D，在堆叠应用处理器（AP）中使用的SAINT L。其中，SAINT S技术已经通过了目前的验证测试。三星很有可能将SAINT这一技术应用于集成高性能芯片所需的存储器和处理器，其中就包括AI芯片。

有消息称，三星内部正考虑将其SAINT 3D封装技术应用于Exynos系列移动处理器上，以进一步提高Exynos处理器的整体性能和生产效率。而在为客户代工方面，还需要与客户进一步测试后才能推出商用服务。此外，还有消息称，三星的SAINT封装技术将为英伟达的Blackwell AI加速器生产关键组件。

目前，台积电在先进封装领域已经具备了较大的技术和生产优势，而英特尔也在美国政府的扶持下不断开拓其3D封装的客户。在竞争如此激烈的当下，三星需要通过加大研发投入力度、提高生产效率，以及优化产品定价等方式来增强自身的竞争力，以赢得市场份额。

GPU的又一挑战者

本报记者 王信豪

就在英伟达财报发布前夕，AI芯片初创公司Groq在社交媒体上引发了广泛讨论。Groq宣称其LPU（语言处理器）的推理性能是英伟达GPU的10倍，而成本仅为其1/10。

英伟达作为人工智能浪潮下的头部企业，近年来不乏“挑战者”对其发起冲击，那么此次LPU的表现如何？

TSP+SRAM的新路径

2月19日，Groq向用户开放了产品体验入口，其产品并非大模型，而类似于大模型加速接口。经由Groq加速推理的开源大模型带给人最直观的感受便是“快”。

根据记者测试，Groq的推理性能达到了每秒270个Token左右，网友测试每秒最高可达500Token，这个速度在ArtificialAnalysis.ai的测试中表现也十分突出。

LPU在LLM和生成式AI上的表现为何快于GPU？

Groq表示，LPU旨在打破LLM的两个瓶颈：计算密度和内存带宽。就LLM而言，LPU的计算能力强于GPU和CPU，这减少了每个单词的计算时间，从而可以更快地生成文本序列。此外，与GPU相比，打破外部内存瓶颈使LPU能够在LLM上提供更好的性能。

在架构方面，Groq使用了TSP（张量流处理）来加速人工智能、机器学习和高性能计算中的复杂工作负载。根据Groq公开技术资料，TSP是一种功能切片的微架构，芯片上具有诸多计算模式被软件预先定义好的功能片，其与数据流的关系如同工厂的流水线。“当数据经过切片时，每个功能单元可以选择性地截取所需数据并获取计算结果，并将结果传递回数据流，原理类似于装配线操作员（功能片）和传送带（数据流）。”Groq公司首席执行官Jonathan Ross比喻道。

TSP的源头是谷歌研发的TPU（张量处理器），值得一提的是，Ross就曾是谷歌TPU研发团队成员之一。

在存储性能方面，LPU另辟蹊径，有别于传统算力芯片对于SK海力士等存储厂商所产HBM（高带宽存储）的依赖，转而使用了易失性存储器SRAM，这也省去了将HBM置于芯片时对台积电CoWoS S封装技术的需求。SRAM通常用于CPU的高速缓存，由于不需要刷新电路来保持数据，因此可提供高带宽和低延迟。

可以说，在张量处理器上的技术积累加上别样的存储器选择，共同造就了这个推理的效率“怪兽”。

实现落地仍有阻碍

尽管在Groq官方口径中，以“快”著称的推理性能确实优于大模型普遍生成内容所使用的GPU，但是从实验室数据到真正流入市场参与竞争，还有许多问题需要解决。

首先，LPU在市场最关心的成本问题上众说纷纭。据Jonathan Ross所说，在大模型推理场景中，Groq LPU芯片的速度比英伟达GPU快10倍，但价格和耗电量都仅为后者的1/10。

看似极高的性价比，实际情况还有待推敲。原阿里技术副总裁贾扬清在社交媒体上算了一笔账，因LPU的内存仅有230MB，在忽略推理时内存损耗的情况下想运行LLaMA2-70b这样的大语言模型可能需要572张LPU，总购卡成本高达1144万美元（按单卡标价2万美元计算）。相比之下，8张英伟达H100的系统在性能上与Groq系统相当，但硬件成本仅为30万美元。

其次是Groq LPU的适用范围能否跟上AI应用的发展速度还是未知数。随着Open AI在2月初发布AI视频生成平台Sora，生成式人工智能走向新的阶段。LPU虽然能实现对Token这一单元的快速处理，但是面对Sora的最小计算单元Patch，其处理效果如何还未可知。有观点认为，LPU在架构上有所创新，但是仅针对特定算法、特定数据结构来设计芯片，在未来频繁改变的AI发展节奏中可能会“昙花一现”。

能否成功挑战英伟达？

再回到“挑战英伟达”的话题上，在Groq所展现出来的解决方案的背后是通用芯片与专用芯片的路径分歧。Groq芯片专注于推理，从测试结果上看能够达到令人满意的“秒回”效果，但是这要依赖对大模型的前置训练环节，换言之，LPU的应用场景搭建，需以至少一个完成且开源的大模型为前提。

英伟达作为GPGPU（通用GPU）的头部生产企业，其A100和H100能够覆盖大模型训练和推理的全部流程，而下一代芯片H200在H100的基础上将存储器HBM进行了一次升级，为的也是提升芯片在推理环节中的效率。在拥有牢固开发者生态的英伟达眼中，强化推理能力也是巩固自身通用GPU市占率的手段。

目前看来，英伟达GPU的交付周期与全球云服务厂商的算力缺口仍存在一定不匹配，英伟达正在积极解决这一问题，与此同时，Groq的LPU能否分得一块蛋糕，还需等待大规模流片之后再看市场反响。

2023年8月14日，Groq宣布三星电子将为自己生产4nm芯片，首批LPU将于2024年下半年量产。Ross表示，在两年后Groq能够部署100万台LPU。

社长：张立 社址：北京市海淀区紫竹院路66号赛迪大厦18层 邮编：100048 每周二、五出版 周二8版 周五8版 零售4.50元 全年定价420元 广告部：010-88558848/8808 发行部：010-88558777 广告许可证：京石工商广登字20170003号 发行单位：中国电子报社 印刷单位及地址：经济日报印刷厂 北京市西城区白纸坊东街2号