2024 年 03 月 14 日

前瞻科技研究

行业深度研究(深度)

证券研究报告

前瞻科技组
分析师：刘道明（执业 S1130520020004）
liudaoming@gjzq.com.cn

电子组
分析师：樊志远（执业 S1130518070003）
fanzhiyuan@gjzq.com.cn

人工智能驱动单芯片 PPA 提升，背部供电将成为行业新趋势

投资逻辑：

半导体行业受 AI 驱动将步入高速增长时代。2023 年，尽管全球半导体销售总额较上一年下降 8.23%，至 5268 亿美元，但自 2023 年 9 月以来同比增速已经回正，2023 年 12 月销售额更是达到 518 亿美元，同比大幅增长 19.12%，显示出行业复苏的明确信号。同时，在最近的国际固态电路会议（ISSCC，2024 年 2 月 18 日至 2024 年 2 月 22 日）上，台积电的高级副总裁张晓强《半导体行业：现状与未来》中也给出了乐观展望：至 2030 年，半导体市场规模有望突破一万亿美元大关，其中，高性能计算，尤其是与人工智能相关的应用，预计将贡献约 40%的收入。AI 相关的技术进步和应用需求，成为行业增长的关键因素，将推动半导体行业步入一个高速增长的新阶段。

人工智能算力需求增长速度远超工艺演进速度。进入大型机器学习模型时代后，训练和推理所需算力翻倍的时间周期分别缩短为 7.4 与 33.8 个月，远快于摩尔定律顶一下晶体管 48.8 个月的翻倍速度。为了满足人工智能爆炸性算力需求，系统摩尔和集群化成为大势所趋。

系统摩尔和集群化面临物理限制，单片 PPA（更高性能，更低功耗，更小面积）仍是提升算力的关键。当前技术及工艺限制下，单芯片性能提升速度不断趋缓，大型集群更多通过系统摩尔及网络并行技术快速提升算力，但边际效应正在递减。通过对特斯拉 DOJO 和英伟达 GH200 的性能比较，我们认为对单芯片单位功耗算力的提升仍是满足算力需求的关键。

背部供电创新性地通过结构改变实现晶体管缩放，是单片 PPA 增长的第二曲线。背部供电能够有效缓解电压降问题，并节省片上空间以容纳更多晶体管，并且这一结构的实现并不依赖于光刻机性能的提升，我们认为在当今摩尔定律放缓的趋势下，背部供电技术将使单片 PPA 重回快速增长，并有望成为未来行业发展中确定性极高的技术方向。

众多大厂已布局，英特尔 PowerVia 技术即将落地，公司有望再度迎来 FinFET 时刻。当前，台积电，三星电子和英特尔均披露已布局背部供电技术，台积电和三星预计分别在 2026 和 2025 年推出该技术，英特尔最为领先，计划在 2024 的 20A 节点中就将其 PowerVia 技术推向市场。经过深入研究后，我们认为英特尔 PowerVia 在性能表现，良率以及客户端工程师开发套件生态等方面都有较为优异的完成度，有望借此技术重新夺回半导体制造领域的领先优势。

投资建议

我们认为，背部供电技术将成为未来半导体制造的新趋势。从技术持有者的角度来看，我们看好英特尔公司通过该技术重新获得半导体制造行业的领先优势。竞争格局方面，背部供电技术不再仅仅是依赖 DUV 或 EUV 的硬件技术来获得领先的晶体管缩放速率，而是通过 DTCO（Design Technology Co-Optimization）从硬件、软件以及设计角度实现晶体管迭代。这一趋势将进一步扩大头部 Fab 的领先优势。综合考虑，从技术持有者的角度，我们建议关注英特尔公司、台积电和三星电子。另一方面，背部供电技术对上游制造设备提出了新的要求，尤其是在混合键合领域。因此，我们建议重点关注应用材料公司和东京电子公司。

风险提示

背部供电技术进展不及预期；宏观经济景气度不及预期。

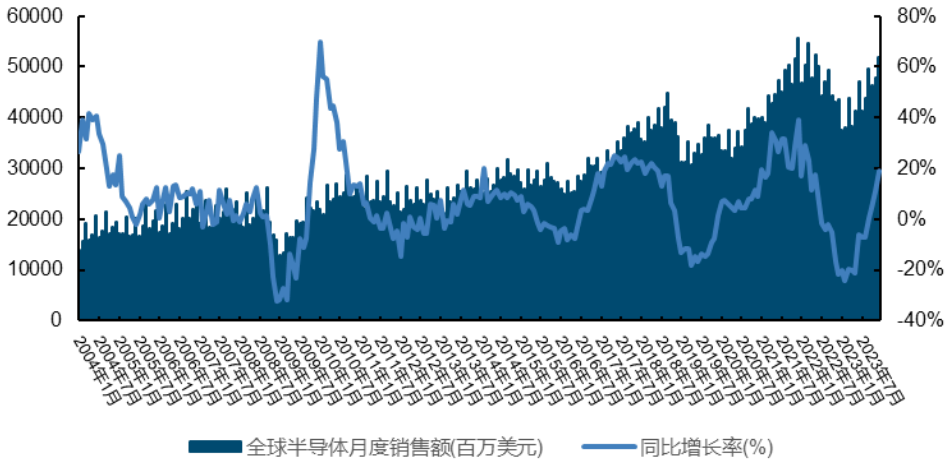
内容目录

图表目录

一、 全球半导体市场进入加速增长期，AI 将是最大驱动因素

对半导体行业而言，经历了两年的下滑之后，最坏的时刻似乎已经过去。世界半导体贸易统计协会(WSTS)的数据显示，行业正在重拾增长势头。2023 年，尽管全球半导体销售总额较上一年下降 8.23%至 5268 亿美元，但自 9 月以来的同比增长率已重新转正并持续增长。到 2023 年 12 月，全球半导体销售额更是同比大幅增长 19.12%至 518 亿美元，显示出行业复苏的明确信号。

图表1：全球半导体市场规模重回增长区间



来源：WSTS，国金证券研究所

展望 2024 年，世界半导体贸易统计协会(WSTS)于 11 月进一步上调了其 5 月的预测，预期 2024 年全球半导体销售额将达到 5880 亿美元，较 5 月预期上涨 124 亿美元，同比 2023 年增长 11.62%。预期上调主要受到逻辑芯片和存储芯片行业的驱动，这两个细分领域预计全年销售额较五月预期分别上涨 64 亿美元和 94 亿美元，较 2023 年同比分别增长 9.6%和 44.8%。逻辑芯片和存储芯片在人工智能领域的广泛应用是支持这一预期增长的主要因素。我们认为，WSTS 对预期的上调调整反映出人工智能将成为 2024 年整体半导体市场的核心驱动力。

图表2：WSTS 五月预期

WSTS Forecast Summary

Spring 2023	Amounts in US$M			Year on Year Growth in %		
	2022	2023	2024	2022	2023	2024
Americas	141,136	128,236	150,989	16.2	-9.1	17.7
Europe	53,853	57,253	61,637	12.8	6.3	7.7
Japan	48,158	48,724	52,534	10.2	1.2	7.8
Asia Pacific	330,937	280,881	310,838	-3.5	-15.1	10.7
Total World - $M	574,084	515,095	575,997	3.3	-10.3	11.8
Discrete Semiconductors	33,993	35,904	38,192	12.0	5.6	6.4
Optoelectronics	43,908	45,949	45,881	1.2	4.6	-0.1
Sensors	21,782	20,410	21,575	13.7	-6.3	5.7
Integrated Circuits	474,402	412,832	470,349	2.5	-13.0	13.9
Analog	88,983	83,907	88,902	20.1	-5.7	6.0
Micro	79,073	71,470	75,855	-1.4	-9.6	6.1
Logic	176,578	173,413	185,266	14.0	-1.8	6.8
Memory	129,767	84,041	120,326	-15.6	-35.2	43.2
Total Products - $M	574,084	515,095	575,997	3.3	-10.3	11.8

Note: Numbers in the table are rounded to whole millions of dollars, which may cause totals by region and totals by product group to differ slightly.

来源：WSTS，国金证券研究所

图表3：WSTS 十一月预期

WSTS Forecast Summary

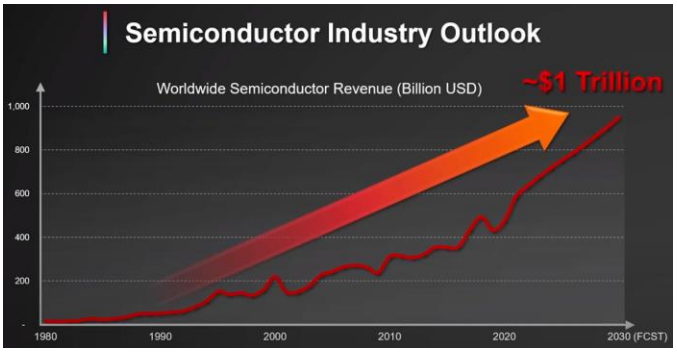
Fall 2023	Amounts in US$M			Year on Year Growth in %		
	2022	2023	2024	2022	2023	2024
Americas	141,136	132,536	162,154	16.2	-6.1	22.3
Europe	53,853	57,048	59,480	12.8	5.9	4.3
Japan	48,158	47,209	49,275	10.2	-2.0	4.4
Asia Pacific	330,937	283,333	317,455	-3.5	-14.4	12.0
Total World - $M	574,084	520,126	588,364	3.3	-9.4	13.1
Discrete Semiconductors	33,993	35,951	37,459	12.0	5.8	4.2
Optoelectronics	43,908	42,583	43,324	1.2	-3.0	1.7
Sensors	21,782	19,417	20,127	13.7	-10.9	3.7
Integrated Circuits	474,402	422,174	487,454	2.5	-11.0	15.5
Analog	88,983	81,051	84,056	20.1	-8.9	3.7
Micro	79,073	76,579	81,937	-1.4	-3.2	7.0
Logic	176,578	174,944	191,693	14.0	-0.9	9.6
Memory	129,767	89,601	129,768	-15.6	-31.0	44.8
Total Products - $M	574,084	520,126	588,364	3.3	-9.4	13.1

Note: Numbers in the table are rounded to whole millions of dollars, which may cause totals by region and totals by product group to differ slightly.

来源：WSTS，国金证券研究所

在最近的国际固态电路会议(ISSCC)上，台积电高级副总裁张晓强对半导体行业的未来作出展望。在其题为《半导体行业：现状与未来》的报告中，张晓强预测到 2030 年，半导体市场规模有望突破 1 万亿美元大关。其中，高性能计算，尤其是与人工智能相关的应用，预计将贡献约 40%的收入。张晓强强调，半导体行业不仅重新进入正增长区间，而且将步入一个高速增长的新阶段，而人工智能相关的技术进步和应用需求将成为推动行业增长的关键因素。

图表4：2030 年全球半导体市场规模有望达到万亿



来源：ISSCC，国金证券研究所

图表5：2030 年 40%的半导体市场收入将来自于 HPC

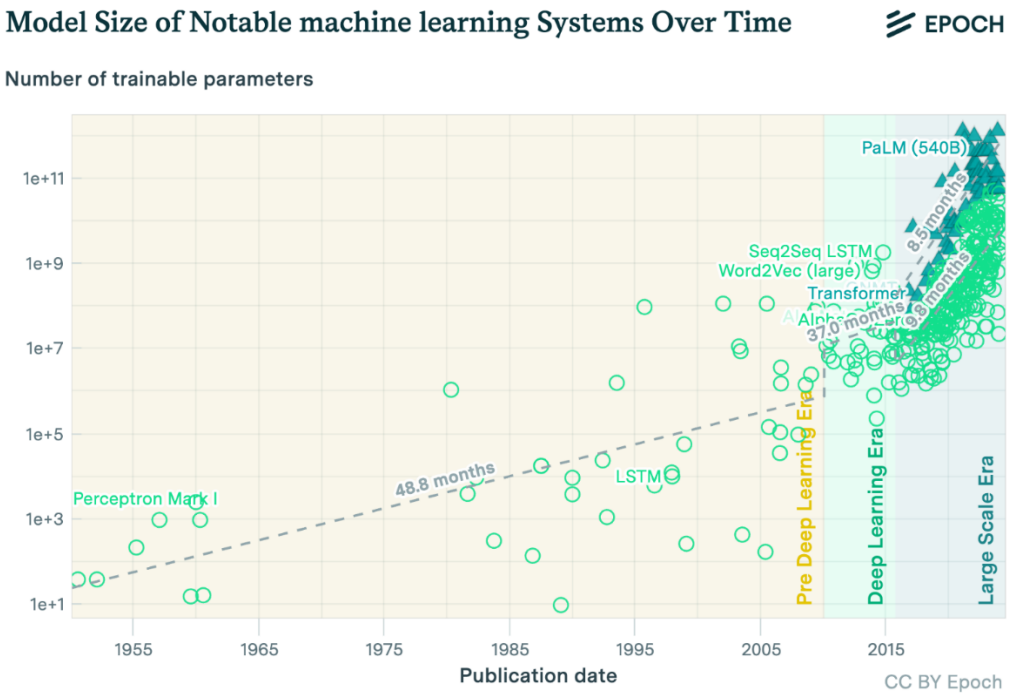
来源：ISSCC，国金证券研究所

二、人工智能相关应用驱动算力需求爆发性增长

1. 大语言模型参数量加速增长，所需算力增长速度远超半导体工艺演进速度

在 2021 年 11 月底，随着 ChatGPT 的发布，大型语言模型（LLMs）在公众视野中取得了突破性进展。这一重大时刻并非偶然发生，而是长期研究和技术进步的积累结果。大型语言模型的核心是大型机器学习模型，其起源可以追溯到 20 世纪 50 年代，当时研究主要集中在规则系统和神经网络的实验上。随着时间的推移，研究者们不断探索新的方法，直到 2010 年代，这些研究成果与崛起的神经网络领域相互交融，为第一个大型语言模型的诞生奠定了坚实的基础。

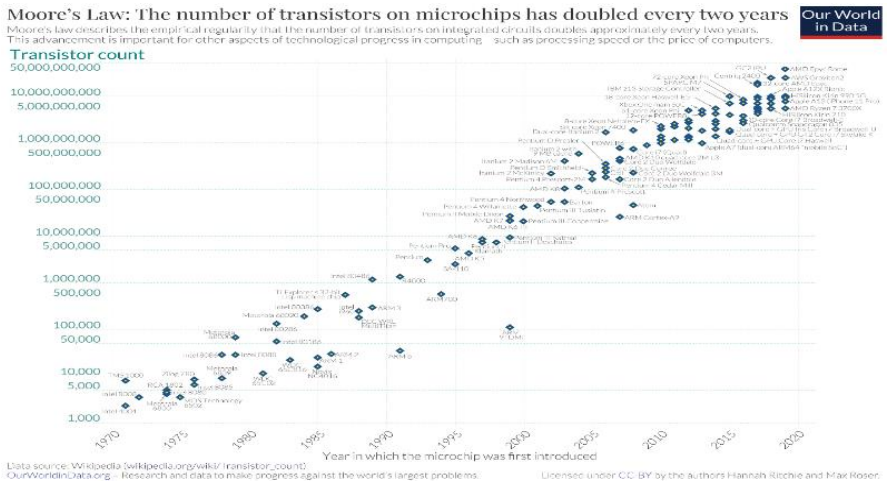
图表6：机器学习模型参数量翻倍所需时间在 2015 年之后显著缩短



来源：EPOCH，国金证券研究所

Google 在 2019 年推出的 BERT 模型是大型语言模型的重要里程碑，它将双向编码、神经网络架构和自监督学习相结合，成为 NLP 任务的标准工具。随后，OpenAI 的 GPT 系列模型（GPT-2 于 2019 年发布，GPT-3 于 2020 年发布）表现出了更强的性能，但与此同时模型参数也大幅增长，GPT-3 的参数量达到了 1750 亿，成为当时最大的语言模型，而最新的 GPT-4 据估计参数量达到了 1.8 万亿。根据 EPOCH 的机器学习模型数据库，当前人工智能模型参数量的增速已经来到了一个前所未有的新高度。

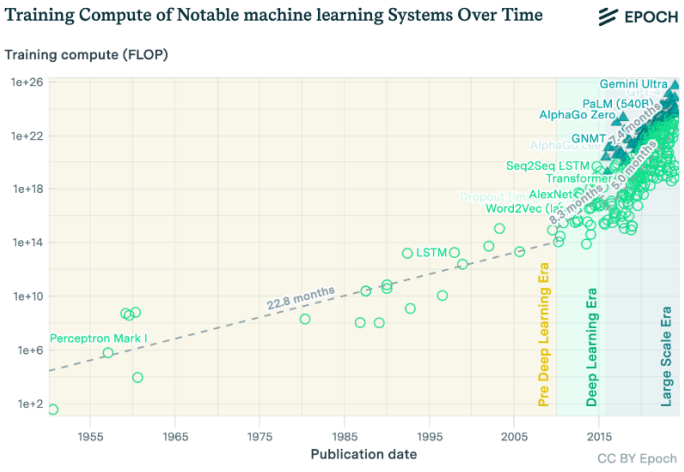
图表7：摩尔定律



来源：Our World in Data，国金证券研究所

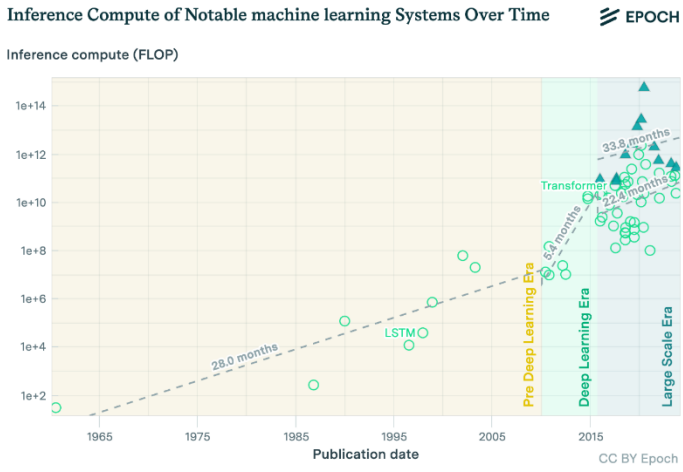
EPOCH 将 AI 计算划分为三个典型阶段，前深度学习时代（Pre Deep Learning Era）、深度学习时代（Deep Learning Era）和超大规模模型时代（Large Scale Era）。在前深度学习时代，机器学习模型参数在 48.8 个月即约两年的时间内翻倍，这一速率和摩尔定律一致。

图表8：大模型时代机器学习模型训练端所需算力需求在 7.4 个月内翻倍



来源：EPOCH，国金证券研究所

图表9：大模型时代机器学习模型推理端所需算力需求在 33.8 个月内翻倍



来源：EPOCH，国金证券研究所

然而在进入深度学习和如今的大型模型时代后，遵循摩尔定律增长的算力开始显著落后于模型所需的算力水平，无论是在训练端和推理端。大型模型时代，训练端和推理端的算力水平分别在仅 7.4 个月和 33.8 个月内就需要翻一倍，远低于晶体管翻倍所需要的时间，这已经足以证明人工智能对计算体系架构的强烈冲击以及对计算架构创新的迫切需求。

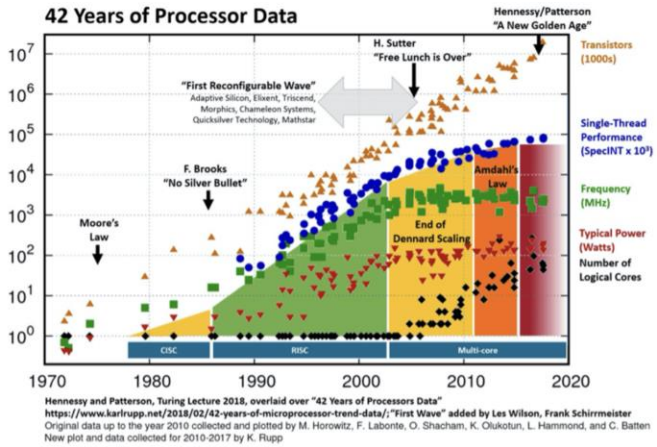
2. 传统摩尔定律受到物理定律限制，系统摩尔势在必行

在半导体行业，摩尔定律一直是推动技术进步的核心原则之一。自 1965 年首次提出以来，它预测了单位面积上晶体管数量的指数增长，大约每两年翻一倍。然而，这种增长并不仅仅体现在晶体管数量的增加上。事实上，芯片性能的提升还依赖于能够在保持功耗不变的同时，将更多的晶体管集成到单位面积上，这一原理被称为登纳德缩放(Dennard Scaling)。

登纳德缩放阐明了一个关键概念：随着晶体管尺寸的缩小，电压和电流也应相应减小，从而维持功耗的稳定。这一原理保证了芯片上可以集成更多的晶体管而不会导致过热问题。然而，这一缩放定律在 2007 年开始失效，并在 2012 年彻底瓦解，标志着计算芯片发展的一个重要瓶颈——功耗墙的出现。

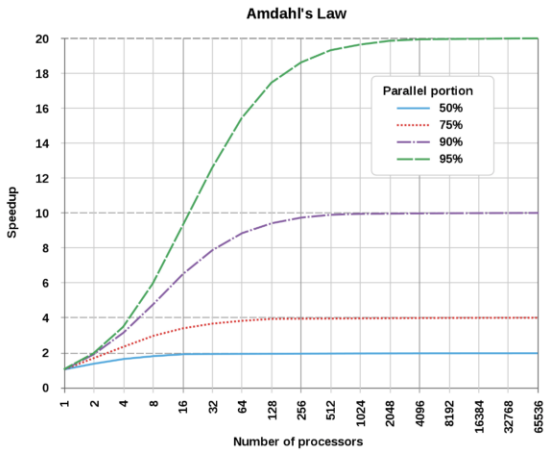
面对登纳德定律的失效，芯片设计转向了通过增加核心数来实现并行计算，以此提升性能。这种多核并行计算在一定程度上延续了性能增长的趋势。然而，这种方法最终受到了阿姆达尔定律（Amdahl's Law）的限制，该定律指出系统性能的提升受限于系统中串行部分的比例。换句话说，仅靠增加晶体管数量（即并行性）是无法无限制提升性能的，这对摩尔定律的持续发展构成了挑战。

图表10：登纳德缩放自 2007 年开始失效



来源：Semiconductor Engineering，国金证券研究所

图表11：系统性能受限于系统中串行部分的比例



来源：VMWare，国金证券研究所

在 2018 年，John Hennessy 和 David Patterson 对未来计算机体系提出了新的构想，开辟了摩尔定律演化的新路径。他们认为，“特定领域的硬件/软件协同设计”将成为半导体领域的核心思想，进而引出了“系统摩尔”的概念。

与传统摩尔定律依赖于晶体管缩放来提升性能不同，系统摩尔着眼于整个系统的设计和集成，强调通过系统级创新实现性能增长。这包括但不限于以下几个方面：

1）异构集成：将处理器、存储器、传感器等不同类型的芯片集成到一个系统中，实现功能的协同工作，提高系统性能和效率。

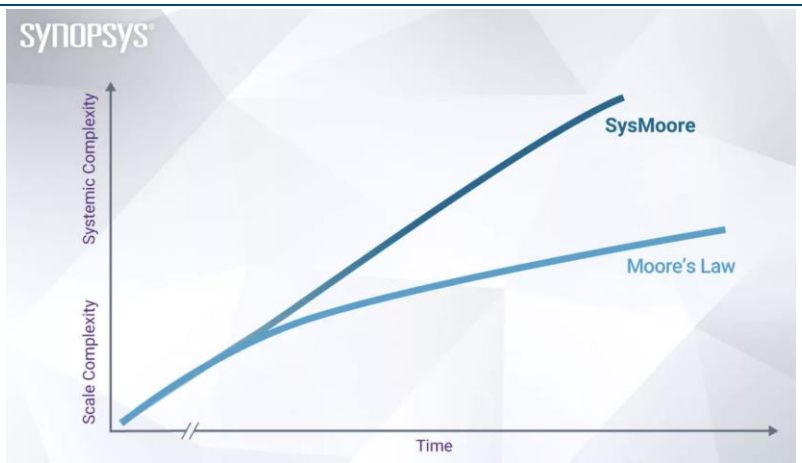
2）先进封装技术：采用 2.5D 和 3D 封装技术，实现不同芯片的紧密集成，缩短信号传输距离，提升速度和能效。

3）系统级优化：包括软硬件协同设计、能源管理和系统架构优化等，充分挖掘系统中各个组件的潜力。

4）域特定架构：针对特定应用领域（如人工智能、高性能计算等）设计专用架构，以提高效率和性能。

系统摩尔旨在传统摩尔定律增长放缓的背景下，通过系统级创新和集成，继续推动性能的增长，满足不断增长的计算需求。

图表12：系统摩尔在传统摩尔定律放缓的背景下通过系统级的创新保持性能增长

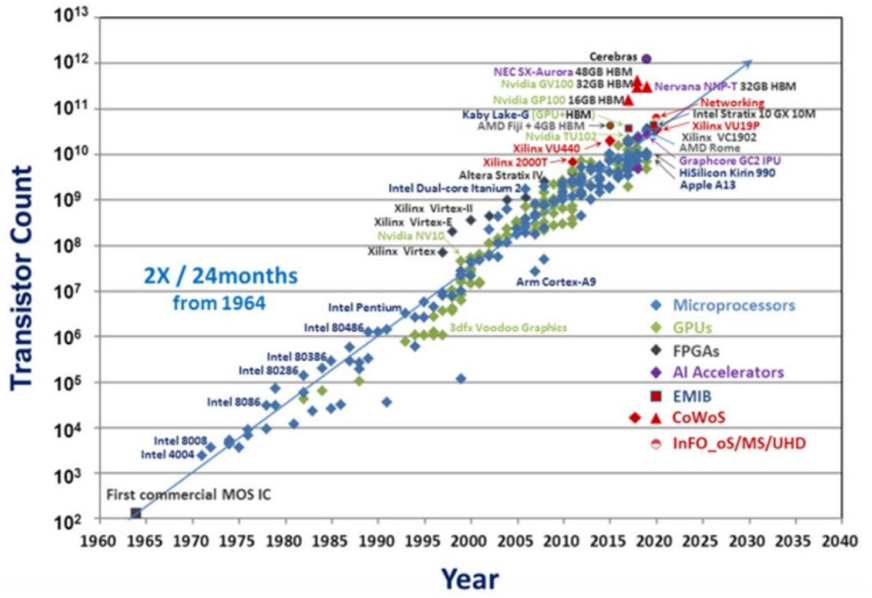


来源：Synopsys，国金证券研究所

3. 系统摩尔性能增速存在理论极限，仍无法满足人工智能巨大算力需求

通过先进封装、特定领域定制化开发（DSA）和高密度互联等技术，系统摩尔能够在两年内使晶体管数量翻一番，达到摩尔定律 1.0 的速度。然而，这一增速上限难以突破，主要受限于先进封装中介层（Interposer）和光照面积限制（Reticle Limit）。当前光照面积约为 858mm2（26mm x 33mm），而先进封装中介层的尺寸为 102mm x 102mm。目前一个封装中介层可以容纳四个光照面积的芯片，预计到 2025 年这一数字将增至六个。尽管基于先进封装的系统摩尔（SysMoore）能够保持摩尔定律的增长速度，但与人工智能所需的爆炸性算力增长相比，系统摩尔仍然难以满足这种巨大的算力需求。

图表13：系统摩尔存在能力上限



来源：《Ultra High Density SoIC with Sub-micron Bond Pitch》，国金证券研究所

4. 网络互联一定程度上突破系统摩尔上限，单片 PPA 仍为性能提升的关键

在传统摩尔定律放缓和系统摩尔面临能力上限的背景下，集群化成为满足大型机器学习模型算力需求的新途径。然而，集群化系统仍然面临阿姆达尔定律所描述的挑战，即并行化带来的算力提升具有边际递减的特性。根据字节跳动近期发布的一篇新论文，该公司已经成功构建了一个由超过一万张卡(12288 张)组成的计算集群，其模型浮点运算利用率(MFU)达到了 55%。MFU 是衡量模型计算效力的重要指标，它的计算方法如下：

$$Model\ \ FLOPs\ \ Utilization = \frac{实际使用的FLOPs}{理论最大FLOPs} \times 100\%$$

图表14：MFU 随卡数增加下降

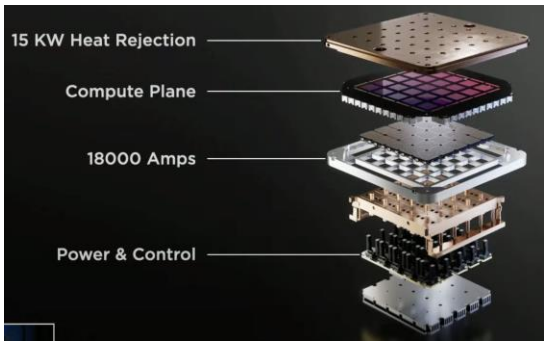
Batch Size	Method	GPUs	Iteration Time (s)	Throughput (tokens/s)	Training Time (days)	MFU	Aggregate PFlops/s
768	Megatron-LM	256	40.0	39.3k	88.35	53.0%	43.3
		512	21.2	74.1k	46.86	49.9%	77.6
		768	15.2	103.8k	33.45	46.7%	111.9
		1024	11.9	132.7k	26.17	44.7%	131.9
	MegaScale	256	32.0	49.0k	70.86	65.3%(**1.23**×)	52.2
		512	16.5	95.1k	36.51	63.5%(**1.27**×)	101.4
		768	11.5	136.7k	25.40	61.3%(**1.31**×)	146.9
		1024	8.9	176.9k	19.62	59.0%(**1.32**×)	188.5
6144	Megatron-LM	3072	29.02	433.6k	8.01	48.7%	466.8
		6144	14.78	851.6k	4.08	47.8%	916.3
		8192	12.24	1027.9k	3.38	43.3%	1106.7
		12288	8.57	1466.8k	2.37	41.2%	1579.5
	MegaScale	3072	23.66	531.9k	6.53	59.1%(**1.21**×)	566.5
		6144	12.21	1030.9k	3.37	57.3%(**1.19**×)	1098.4
		8192	9.56	1315.6k	2.64	54.9%(**1.26**×)	1400.6
		12288	6.34	1984.0k	**1.75**	55.2%(**1.34**×)	2166.3

来源：《MegaScale: Scaling Large Language Model Training to More Than 10,000 GPUs》，国金证券研究所

根据论文中披露的数据，随着 GPU 数量的增加，模型浮点运算利用率（MFU）逐渐下降，

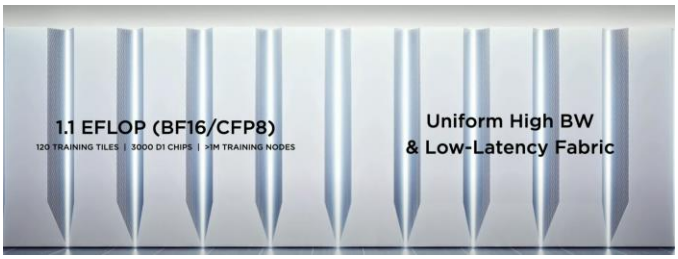
而且总体算力并不是随着 GPU 数量线性增长，而是呈现边际效益递减的趋势。这正是计算集群所面临的阿姆达尔定律所描述的困境。

图表15：DOJO Training Tile



来源：Tesla，国金证券研究所

图表16：DOJO ExaPOD



来源：Tesla，国金证券研究所

除了模型浮点运算利用率（MFU）和算力角度的瓶颈，大型算力集群还面临着电力、散热以及占地面积等物理层面的难题。以特斯拉的 DOJO 计算集群为例，其基础单位是 D1 芯片，单芯片能提供 362 TFLOPS（BF16/CFP8）的算力，但热设计功耗高达 400 瓦。D1 芯片的上层组成单元，即 Training Tile，由 25 个 D1 芯片组成，在不考虑能量耗散的情况下，其热设计功耗可达 10000 瓦，实际工作功耗常超过热设计功耗，尤其是在高性能计算场景中。DOJO 集群的最终形态，DOJO ExaPOD，由 120 个 Training Tile 组成，能提供 1EFLOPs 级别的算力，热设计功耗达到惊人的 120 万瓦。在一次测试中，DOJO 系统的工作功耗达到了 2.3Megawatts，影响到了圣何塞当地的电站，导致实验不得不终止。

这么高的功耗反映出大规模集群对客观物理定律的挑战，以及 DOJO 自身节点性能的不足。相比之下，英伟达的 DGX GH200 由 256 块 450-1000 瓦的 Grace Hopper Superchip 构成，提供同样级别的算力，但在考虑 1000 瓦的极端情况下，其整体热设计功耗为 25.6 万瓦，仅为 DOJO ExaPOD 的 20%左右。这表明，在相同算力下，GH200 能够降低约 80%的功耗，归功于 Grace Hopper Superchip 在节点端单位功耗算力上的优势

我们认为，以上案例充分说明了节点性能的提升仍旧是人工智能和高性能计算向前发展的关键，行业对于高性能，低功耗计算的核心需求也就是单芯片 PPA 从未改变，并且随着算力需求的持续增长，对于单片性能 PPA 的提升的追求将更为极致。

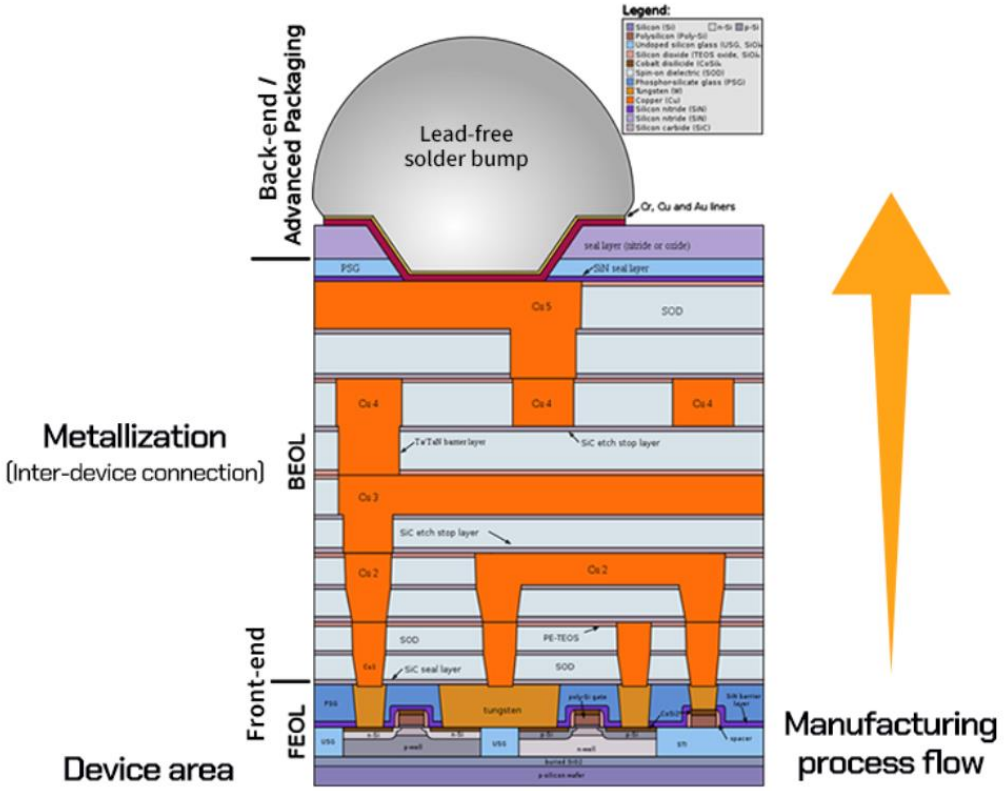
三、背部供电，单片 PPA 的新路径

1. 认识传统半导体

要充分理解背部供电技术的价值主张，首先需要了解传统芯片结构。半导体制造过程分为前端制程（FEOL，Front-End of Line）和后端制程（BEOL，Back-End of Line）。前端制程涵盖了所有在硅片上形成晶体管及其他活性或被动元件的步骤，包括掺杂、氧化、沉积、光刻和蚀刻等过程，这是芯片制造中最关键的部分。在 FEOL 阶段，晶体管作为计算和存储的基本单元被构建在硅片上，其性能直接影响到最终芯片的性能。后端制程则涉及将这些晶体管和其他元件通过金属互连层连接起来，构建成完整的电路。BEOL 包括多层金属导线的沉积、绝缘层的形成、以及通过光刻和蚀刻过程形成导线和绝缘层中的通孔，以实现不同层之间的电连接。BEOL 阶段是确保电信号能够在芯片内部有效传输的关键，同时也是电源和地线分布的主要阶段。

在传统的芯片结构中，电源和信号都是在 BEOL 阶段通过同一侧的金属层进行分配和传输的。这种设计容易导致信号干扰和电源降压（IR Drop）问题，特别是随着芯片性能的提升和制程的缩小，这些问题变得更加显著。电源线路的电阻和电感效应会导致电源到达晶体管时的电压降低，从而影响芯片的性能和能效。

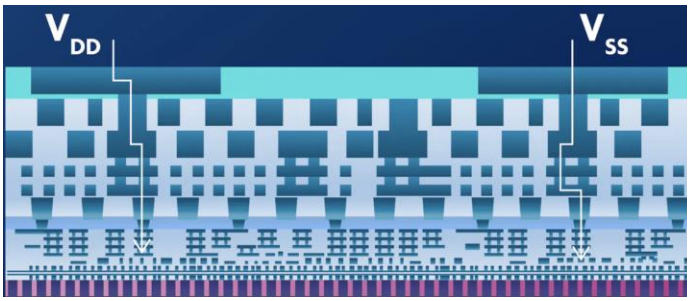
图表17：半导体制造分为前端流程和后端流程



来源：SK Hynix，国金证券研究所

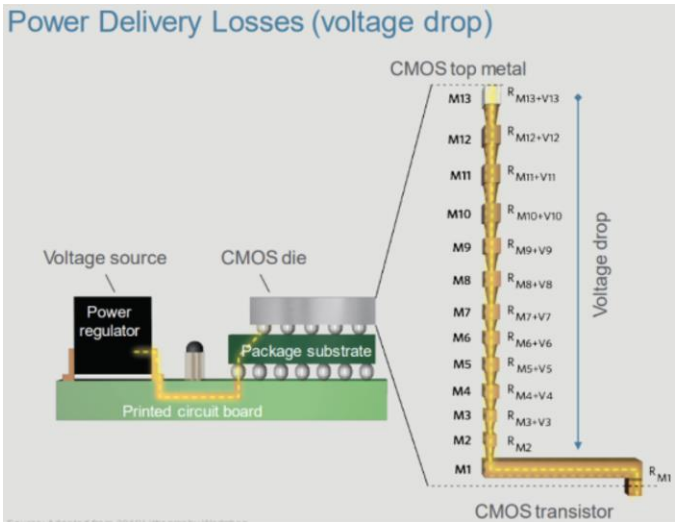
2. 传统半导体结构使晶体管缩放面临瓶颈，背部供电应运而生

图表18：传统半导体采用正面供电



来源：imec，国金证券研究所

图表19：正面供电结构带来电压降（IR Drop）问题



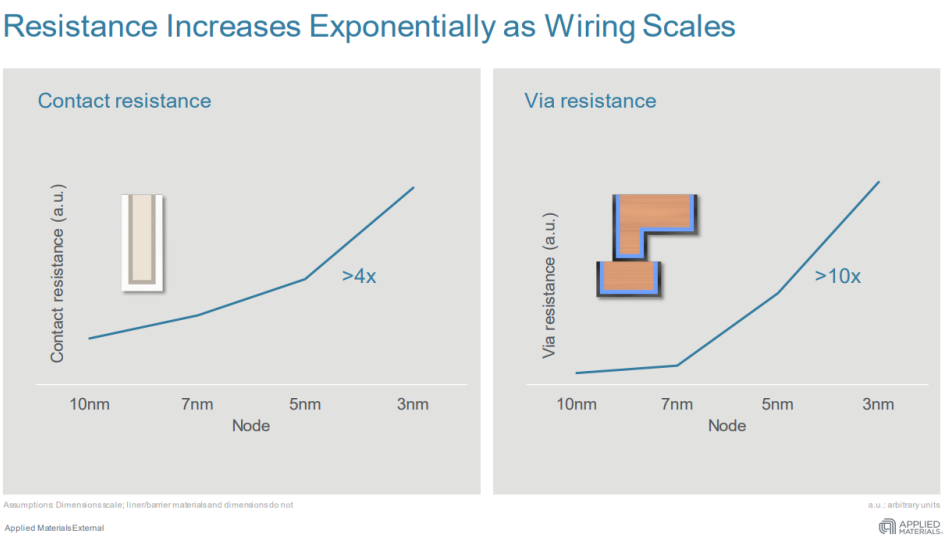
来源：Applied Materials，国金证券研究所

这一多层的结构带来了两大影响芯片性能的问题，并且随着晶体管尺寸逐渐缩小日益显著：

1） IR Drop/Droop：此处的'IR'中的'I'即为电流，'R'即为电阻，'IR'即为电压，IR Drop（电压降）问题是指在半导体芯片的电源分配网络中，由于电流通过导线时导线本身的电阻作用，导致电源到达负载（如晶体管）时电压降低的现象。这个问题在集成电路（IC）设计中非常关键，因为它直接影响到芯片的性能和可靠性。如果 IR Drop 过大，可能会导致芯片无法正常工作，特别是在对电压敏感的模拟电路或是高速数字电

路中。通常来说，芯片工程师会保证电压降的幅度小于 10%。导致电压降的原因主要有以下几点：

图表20：通孔电阻随着制程下降加速增大



来源：Applied Materials，国金证券研究所

a） 通孔材料所导致的通孔电阻(Via resistance)激增：历史上，由于铜的电阻较低，所以一直使用铜来制造通过孔。但是，铜需要一个扩散屏障，比如使用钽氮化物（TaN）作为屏障材料，这带来了两个问题。首先，由于屏障材料占据了一定的空间，它减少了通过孔中铜的横截面积。其次，屏障材料位于通过孔的底部，电流必须通过这层屏障材料流动。由于屏障材料的电阻比铜高，这就导致了通过孔的电阻增加。简而言之，虽然铜是制造通过孔的理想材料，但是必须使用的屏障材料在一定程度上增加了通过孔的电阻，这是半导体制造过程中需要解决的技术难题。除此之外，随着晶体管逐渐缩小，尤其是在极小的尺寸下，接触电阻和通孔电阻都以更快地速度增长，进一步加剧了电压降这一问题。

b） 多层金属化导致的高电阻：随着集成度的提高，芯片上集成了越来越多的功能和逻辑单元。为了在有限的空间内实现这些复杂的电路设计，需要多个金属层来有效地布线和互连这些单元。多个金属层同时也是实现电源和信号路径相分离的必要条件。就金属层本身的电阻而言，仅 M0 一层就达到了 900Ω/μm，十余层叠加导致电压传输路径上电阻非常高。

2） 电源和信号电路的互连问题，信号布线和电源分配网络的对电路的要求是不同的。电源需要低电阻，高电容，因为它需要传输大电流，而对于普通信号来说它可以承受一定的电阻，但他需要低电容。解决方案是区分电源和信号布线，代价是更复杂的工艺，使晶体管的缩放愈发困难。

图表21：电源和信号网络对于电路硬件的要求不同

Signal wiring	Power Delivery Network
low Capacitance	high capacitance
want small x-section	want large x-section
can tolerate some R (often small current, or use My)	very (!) low Resistance
	Electromigration concern
Resistance at drain of the FET ➔ smaller impact on perf	Resistance at source-side of the FET ➔ huge impact on perf

来源：Cadence，国金证券研究所

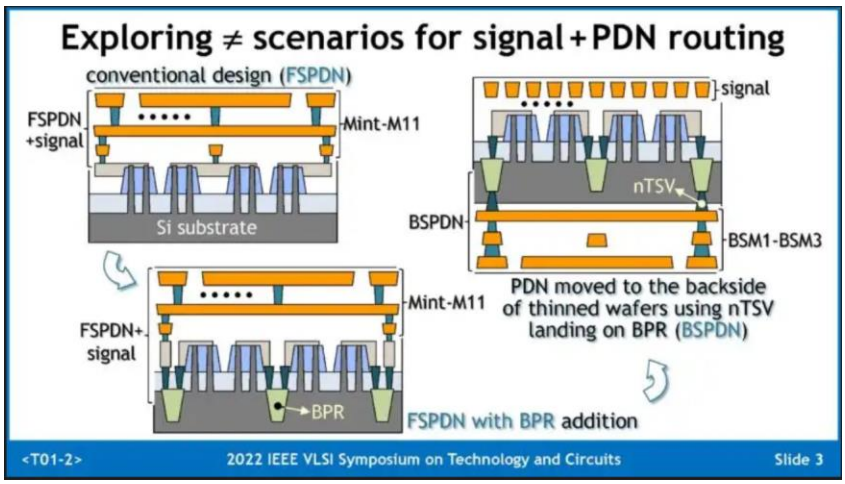
为了应对以上问题，背部供电应运而生，背部供电的理念是通过将电源分配网络移到硅片

的背面来解决这一问题，从而实现了电源和信号路径的物理分离。这样不仅可以减少信号干扰，还能通过使用更宽、电阻更低的金属线来改善电源传输效率，从而减轻 IR Drop 问题。通过这种方式，背部供电不仅优化了电源管理，还为未来芯片设计提供了更大的灵活性和性能提升的可能性，使得芯片设计能够更好地满足高性能计算和低功耗的需求。

3. 初代背部供电网络能够有效降低电压降低幅度，并帮助提升单位面积晶体管数量

背部供电(BSPDN，Backside Power Delivery Network)的概念最早由 imec 于 2018 年 IEEE IEDM 会议上提出并定义了初代背部供电的物理结构背部埋藏电路 BSBPR(Backside Buried Power Rail)。埋藏电路是一种金属线构造，埋藏在晶体管下方——部分位于硅基底内，部分位于浅沟槽隔离氧化物内。它承担了传统上在 BEOL（后端制程）标准单元层级实现的 VDD 和 VSS 电源轨的角色。

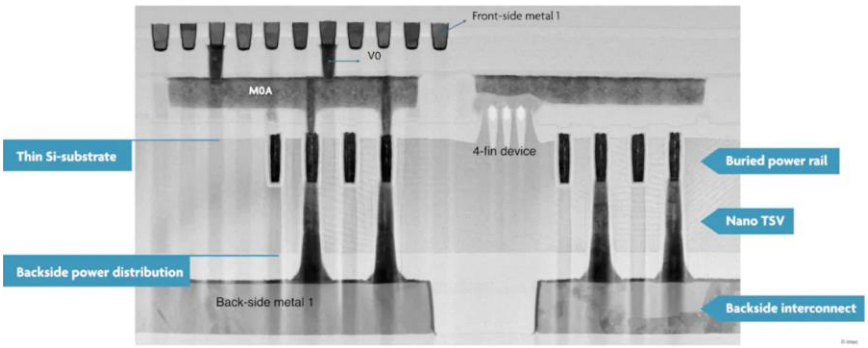
图表22：BSBPR 结构示意图



来源：IEEE，国金证券研究所

背部埋藏电路实际上正面供电埋藏电路的迭代版本，正面埋藏电路最初的目的是为了解决在 M2 金属层上电源线路占据过多面积的问题，在一个标准逻辑单元的上下分别连接着电源正极线和地线，传统半导体结构中为了缓解电压降问题，在 M2 层级上的电源线路通常比其他导线更宽，然而更宽的电源导线同时也占据了更多的面积限制了晶体管缩放，正面埋藏电路就是为了解决这个问题，但电源仍旧从晶圆的正面，即晶体管上方沿着更高级别的金属层向更低级别的金属层传输。背部埋藏电路保留了埋藏电路，但是将整个供电网络移到了晶圆背面，背部供电网络通过 Nano TSV 连接到埋藏电路，实现了供电网络和信号网络的分离。

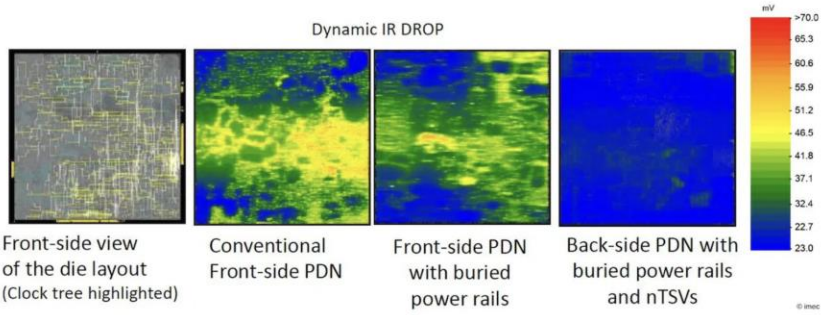
图表23：BSBPR 的透射电镜图像



来源：imec，国金证券研究所

能效方面，目前已有众多研究表明，BSBPR 能够有效降低电压降低幅度。imec 在 2019 年的 IEDM 会议上展示了他们和 ARM 合作开展的一项通过模拟比较传统正面供电，正面埋藏电路和背面埋藏电路 CPU 电压降低幅度的研究，模拟结果显示背面埋藏电路相较正面埋藏电路能够将电压下降的幅度缩减 7 倍。

图表24：BSBPR 将电压下降幅度缩减 7 倍



来源：imec，国金证券研究所

2022 年一项德州大学奥斯丁分校的研究对 5nm 以下 FinFET 晶体管在三种供电系统下的功耗表现进行了实证研究，结果显示正面埋藏电路和背部埋藏电路相较于普通正面供电系统，功耗分别为 1.1 和 0.92 倍，电压降低幅度分别为 0.75 和 0.15 倍。

图表25：各供电方案关键指标比较

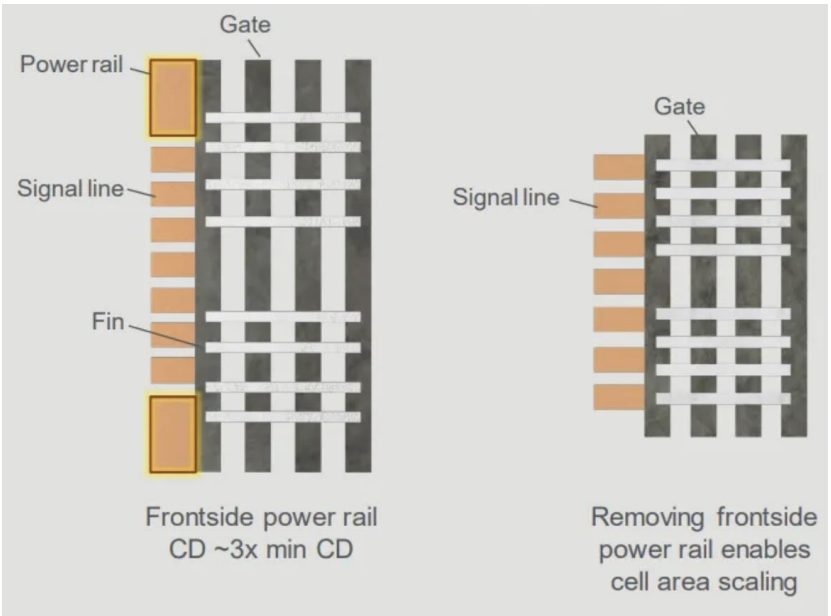
COMPARISON OF IMPORTANT DESIGN METRICS

PDN	Area	Frequency	Power	IR drop	Power supply noise	Local power grid resistance
FS	1x	1x	1x	1x	1x	1x
FSBPR	1x	1x	1.1x	0.75x	0.85x	0.25x
BSBPR	1x	1x	0.92x	0.15x	0.7x	0.02x

来源：《A Holistic Evaluation of Buried Power Rails and Back-Side Power for Sub-5 nm Technology Nodes》，国金证券研究所

除此之外，背部供电可以将电源轨道移到晶圆背面以放置更多的晶体管，通常来说传统正面电源轨道的尺寸是晶体管网络关键尺寸(Critical Dimension)的三倍，因此将电源轨道迁移到晶圆背面对晶体管缩放而言意义重大，一般来说可以提供额外 15～20%的晶体管面积以放置更多的晶体管。

图表26：通过背部供电可以节省片上空间以容纳更多晶体管



来源：Applied Materials，国金证券研究所

四、众大厂布局，PowerVia 先声夺人

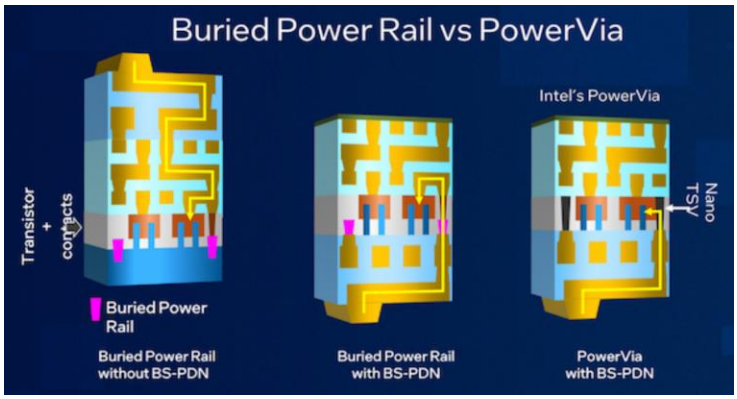
在 imec 于 2018 年的 IEDM 会议上首次提出了背部供电方案后，众多大厂跟进：

1） 三星电子在 2023 年 VLSI 研讨会上披露了其背部供电网络(BSPDN)技术的相关参数，该方法成功将所需的晶圆面积减少了 14.8%，为芯片腾出更多空间容纳更多晶体管，提高整体性能；同时布线长度减少了 9.2%，有效降低电阻，使更多电流能够通过，从而降低功耗，优化功率传输效率。三星电子首席技术官 Jung Ki-tae Jung 当时表示计划在 2027 年将 BSPDN 应用于 1.4nm 工艺。不过，根据韩国业内人士于 2 月 28 日的最新透露，三星电子证实其研发中的背部供电技术指标超出预期，由于技术进展顺利，三星电子预计将提前将背部供电技术推向市场，最快在 2025 年就应用于 2nm 工艺节点。

2） 台积电采取了一种相对保守的背部供电技术路线，采用了低复杂度的埋藏电路设计，可利用现有工具实现，技术成熟度较高。在 2023 年台积电技术研讨会上，台积电透露将在 N2P 制程节点引入背部供电技术，有望缓解电压降问题，改善信号完整性，使性能提升 10%-12%，同时将逻辑电路面积缩减 10%-15%，台积电计划于 2026 年推出 N2P 制程，届时将应用上述背部供电技术。

3） 英特尔是目前对背部供电押注最大的厂商，英特尔于 2022 年 2 月的 IEEE 国际固态会议(ISSCC)上首次提出了他们的背部供电方案，该方案被称为 PowerVia。

图表27：PowerVia 和 BPR 结构对比



来源：Intel，国金证券研究所

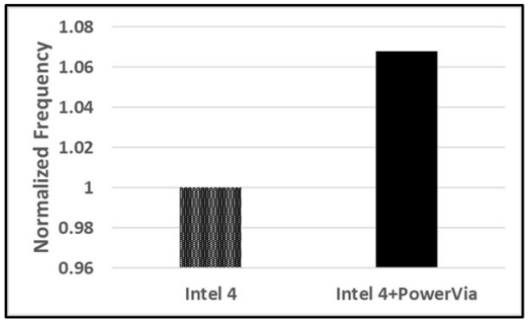
相较于 BSBPR，PowerVia 的背部供电网络通过和晶体管同一层级的 nTSV 直接连接到晶体管的源/漏极上，进一步释放了晶体管正面 M0 层上原先被电源线路占据的空间。根据英特尔于 2023 年 VLSI 论坛发布的一篇研究，英特尔构建了一块代号为 Blue Sky Creek 的逻辑芯片以测试 PowerVia 技术对 Intel4 工艺芯片性能的提升。得益于我们先前讨论的 PowerVia 可以进一步释放 M0 层上的空间，Intel4+PowerVia 样本的 M0 pitch 从 Intel4 样本的 30nm 提升至 36nm，等同于 Intel7 工艺的尺寸，这将帮助降低晶圆正面 Fabrication 过程的工艺复杂度。性能方面，Intel4+PowerVia 样本相较 Intel4 样本提升了 6%的核心频率，效率核心(E-Core)实现了超过了 90%的单元利用率(cell utilization)，表明芯片上用于放置逻辑单元的面积与总可用面积的比值大于 90%，相较一般高性能 CPU 的 60～80%的单元利用率取得了显著的提升。

图表28：PowerVia 降低前段工艺难度

	Intel 4	Intel 4 + PowerVia
Contacted Poly Pitch (nm)	50	50
Fin pitch (nm)	30	30
M0 pitch (nm)	30	36
#front-side layers	15+RDL	14
#back-side layers	n/a	4+RDL
HP library height (nm)	240	210

来源：Intel，国金证券研究所

图表29：PowerVia 提升了 6%的核心效率



来源：Intel，国金证券研究所

根据英特尔 PowerVia 背面供电技术的官方介绍，英特尔将在 Intel 20A(2nm)将是英特尔首个采用 PowerVia 背面供电技术及 RibbonFET 全环绕栅极晶体管的节点，预计将于 2024 年上半年实现生产准备就绪，在时间线上将领先台积电和三星至少一到两年。

五、背部供电工艺流程

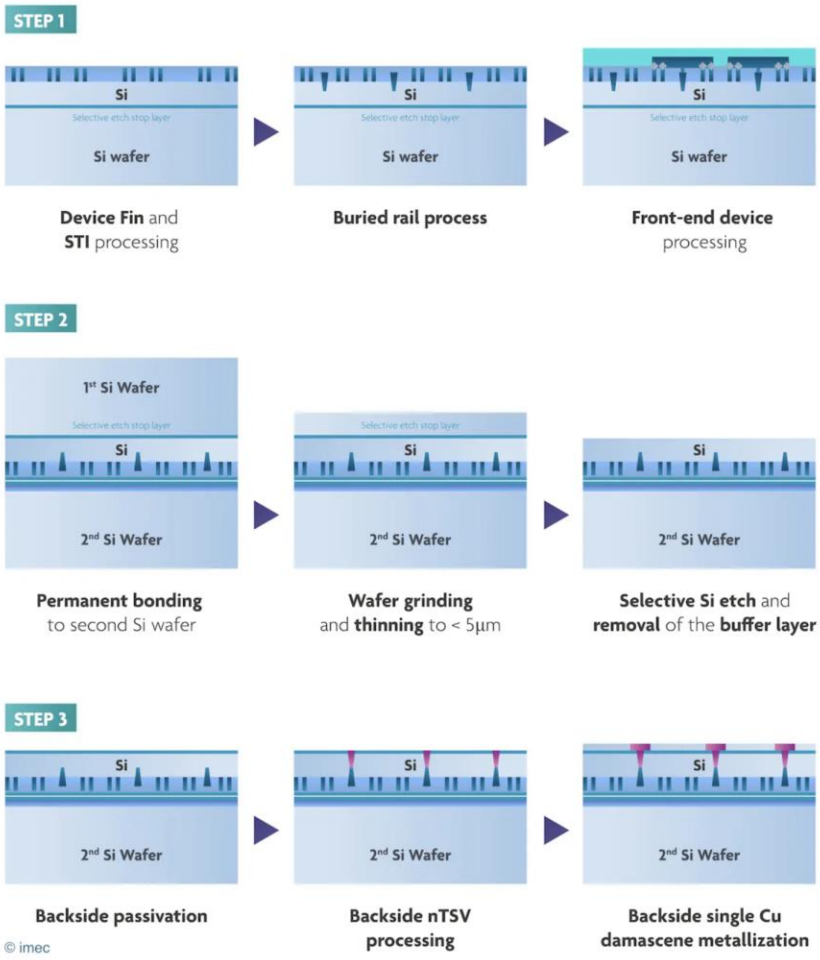
imec 为其 BSBPR 提供了非常详细的工艺流程，BSBPR 的工艺流程大致可以分为以下三步：

1） 正面晶体管和埋藏电路制造：该工艺流程从在 300mm Si 晶圆上生长 SiGe 层开始。SiGe 层将在第二步中作为蚀刻停止层，用于终止晶圆减薄流程。接下来，在 SiGe 层上生长一层薄的 Si 覆盖层：这是制造晶体管和埋藏电源导轨的起点。埋藏电源导轨的位置由浅沟槽隔离带决定。在 Si 覆盖层中蚀刻的沟槽被氧化物衬里和金属填充，例如 W 或 Ru。得到的埋藏导轨通常宽约 30nm，间距约 100nm。晶体管(如 FinFET)在 BPR 制造完成后才进行刻蚀，BPR 通过 VBPR 和 M0 层上的导线连接到晶体管的源/漏极。在晶体管和 BPR 都制造完成后，再实施铜金属化以实现线路互连。

2） 晶圆键合(Wafer-to-Wafer Bonding)和晶圆减薄：前端制程完成的晶圆被翻转后通过介电融合键合连接到一块搭载晶圆上(Dummy Silicon)以实现晶圆背面的固定。随后对被翻转的晶圆的背面进行减薄直至触及到第一步中生长得到的 SiGe 层，整个过程通过依次进行背面研磨、化学机械抛光以及干法和湿法刻蚀的步骤来实现。

3） 制造 nTSV 并连接到 BPR：首先在减薄的晶圆背面上沉积一层钝化层，使用硅衬底光刻(Through-Si Alignment Lithography)在硅层上刻蚀出 nTSV 并触及到 BPR。随后在 TSV 的内壁上沉积一层氧化物作为衬里(oxide liner)，然后再填充金属钨(W)作为导电材料。该流程通过加工一个或多个背面金属层完成，通过 nTSVs 将晶圆的背部供电系统与正面的 BPR 连接起来。

图表30：背部供电工艺流程



来源：imec，国金证券研究所

英特尔 PowerVia 的工艺和 imec 的 BSBPR 大致流程相同，同样也需要在晶圆正面制造完成后将其翻转键合到一个搭载晶圆(Carrier Wafer)上，随后再通过光刻流程在晶圆背面刻蚀得到 nTSV 并将其与 BPR 连接。

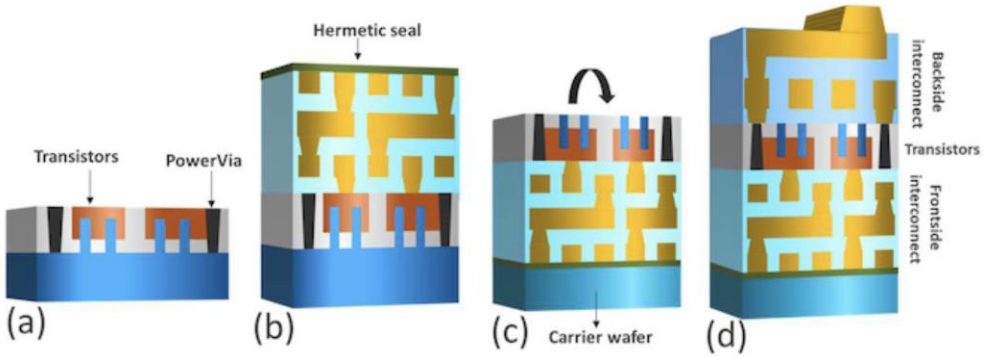
在整个工艺流程有几个关键步骤可能会面临一定的工艺难度：

1) 晶圆减薄后硅金属层厚度均一性：为了制造 nTSV 并尽量缩短他们的长度以最小化电阻率，需要将晶圆背面减薄至仅保留厚度为几百纳米的硅，在这个厚度尺寸下，硅金属层的厚度不均一将会非常显著。

2) 晶圆间键合的准确性：由于晶圆翻转使 BPR 的位置在 xy 平面上产生了位移，对后续 nTSV 光刻位置的精度提出了更高的要求。传统的光刻校准设备难以直接针对由于晶圆翻转导致的误差，目前可以通过更为先进的晶圆间键合设备和光刻设备来缩小校准误差。

3) 硅衬底光刻(Through-Si Alignment Lithography)：nTSV 在翻转晶圆的背面通过硅衬底光刻生成，硅衬底光刻是一种利用硅衬底作为光刻掩模的新型光刻工艺。与传统的基于掩模的光刻工艺相比，硅衬底光刻利用硅衬底的精确晶体结构来实现更高的对准精度，从而提高芯片的良率。但目前，TSAL 工艺仍处于研发阶段，但它有望成为下一代光刻技术。随着技术的进步，TSAL 工艺有望在精度、尺寸和成本方面超越传统的光刻工艺。

图表31：PowerVia 工艺流程

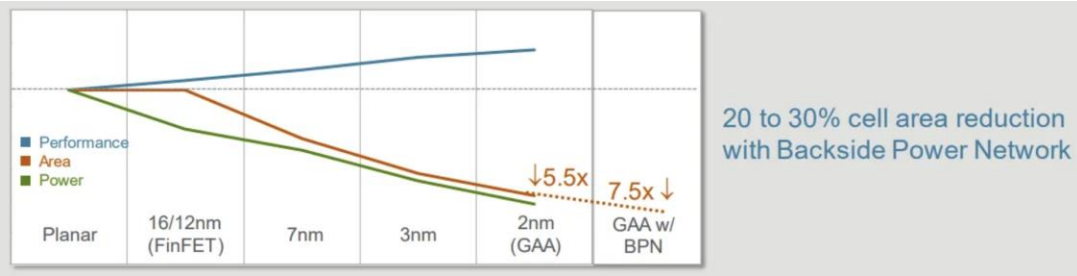


来源：anandtech，国金证券研究所

六、PowerVia 有望成就 Intel 的下一个 FinFET 时刻

我们认为背部供电方案是一个非常具有开创性的实现晶体管缩放的工艺(transistors scaling)，通过背部供电实现的晶体管缩放不依赖于 EUV 技术，在摩尔定律逐渐收到物理极限制约的当下，晶体管缩放不受 EUV 限制的意义不仅仅在于 Fab 厂可以节省设备迭代的费用，而在于 Fab 厂可以通过新的工艺重新实现更快速的晶体管缩放。根据应用材料的估计，PowerVia 的实现等效于两代晶体管缩放的效果。

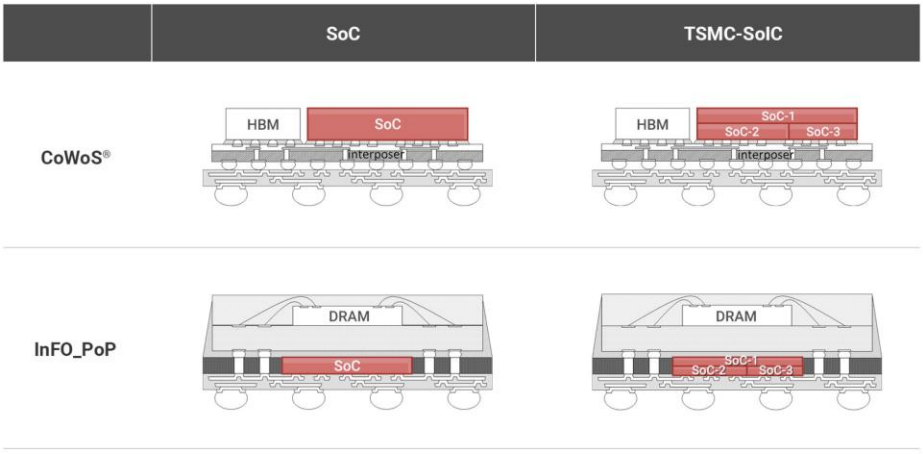
图表32：PowerVia 有望带来等效于两代晶体管缩放的 PPA 提升



来源：应用材料，国金证券研究所

除了晶体管 PPA 性能角度之外，理解背部供电的另一个角度是结构上的变化能够帮助实现更先进的封装。imec 认为得益于信号和电源分离，z 轴上的 SoC 叠加，比方说逻辑 die 和存储 die 叠加，该连接方式可以通过将逻辑 die 的正面和存储 die 的正面键合来实现，这和现在 DRAM 叠加形成的 HBM 是异曲同工。TSMC 的 SoIC 采用的也是类似的思路，我们认为这将是未来先进封装的大趋势，而背部供电是通向这一封装形式的关键工艺。

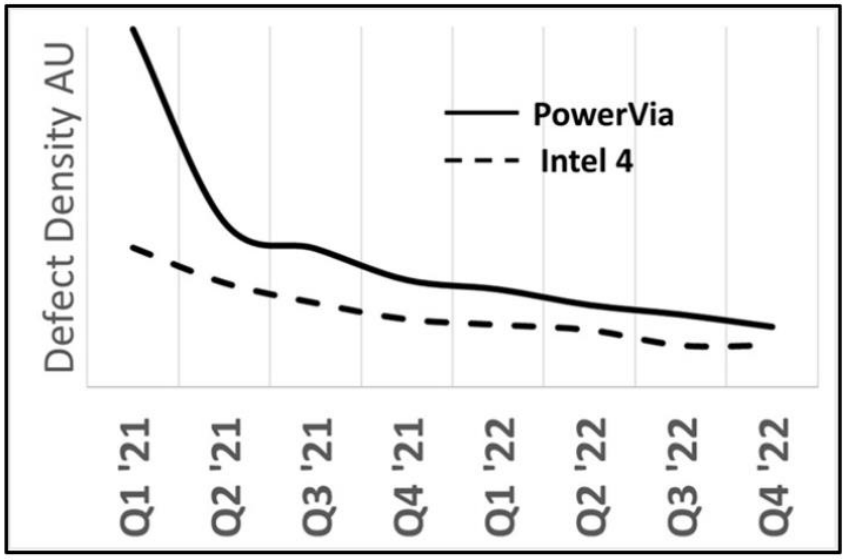
图表33：SoIC 实现 SoC 纵向堆叠



来源：TSMC，国金证券研究所

回到英特尔本身而言，尽管背部供电是一个崭新的技术，已有迹象表面英特尔已经成功完成开发了该工艺。在英特尔 2023 年 VLSI 论坛发布的研究中，结果显示英特尔已经将 PowerVia 的良率大幅提升，在 2022Q4 时，PowerVia 的良率落后 Intel4 工艺良率约两个季度。

图表34：PowerVia 良率接近 Intel4



来源：Intel，国金证券研究所

另外，在 2023 年六月的一场访谈(The Futurum Group，2023)中，英特尔的两位研发人员表示已经成功地解决了几个关键问题：

1） 芯片测试：通常来说，芯片测试通过芯片的正面进行，英特尔的已经成功地在 PowerVia 这个新的结构的背面上实现了精准的芯片测试(通过预埋坏点来验证测试的准确性)。

2） 热学相关测试和验证：背部供电一直被业界认为会存在较大的散热问题，因为在晶圆两侧都有电路。英特尔表示他们已经成功对 PowerVia 结构完成了热学建模并开发了相应的解决方案以供客户使用。

3） EDA 开发套件配套：英特尔已经和多家 EDA 厂商合作，完成开发了 20A、18A 以及 PowerVia 相关工艺的开发套件开发，当前的功能将能够满足芯片工程师的设计需求。

从下游应用而言，背部供电满足当下 AI，图像渲染对芯片高性能、低功耗的要求，在后摩尔时代，背部供电有望成为半导体制造新趋势，英特尔作为该领域的领导者，有望深度受益。

七、投资建议

我们认为，背部供电技术将成为未来半导体制造的新趋势。从技术持有者的角度来看，我们看好英特尔公司通过该技术重新获得半导体制造行业的领先优势。竞争格局方面，背部供电技术不再仅仅是依赖 DUV 或 EUV 的硬件技术来获得领先的晶体管缩放速率，而是通过 DTCO(Design Technology Co-Optimization)从硬件、软件以及设计角度实现晶体管迭代。这一趋势将进一步扩大头部 Fab 的领先优势。综合考虑，从技术持有者的角度，我们建议关注英特尔公司、台积电和三星电子。另一方面，背部供电技术对上游制造设备提出了新的要求，尤其是在混合键合领域。因此，我们建议重点关注应用材料公司和东京电子公司。

风险提示

1） **背部供电技术进度不及预期：**背部供电工艺流程仍面对诸如混合键合，晶圆减薄和硅衬底光刻等工艺难题，若背部供电技术因为以上原因晚于预期推向市场，或性能不及预期，则对产业链相关公司会造成一定不利影响。

2） **宏观经济景气度不及预期：**半导体先进制程推出早期往往需要晶圆厂和终端厂商通力合作以实现规模化技术验证和迭代，若宏观经济景气度下行，终端厂商对低良率产品容忍度下降，则会影响背部供电技术发展速率。

特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

上海
电话：021-80234211
邮箱：researchsh@gjzq.com.cn
邮编：201204
地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 5 楼

北京
电话：010-85950438
邮箱：researchbj@gjzq.com.cn
邮编：100005
地址：北京市东城区建内大街 26 号新闻大厦 8 层南侧

深圳
电话：0755-86695353
邮箱：researchsz@gjzq.com.cn
邮编：518000
地址：深圳市福田区金田路 2028 号皇岗商务中心 18 楼 1806

【小程序】
国金证券研究服务

【公众号】
国金证券研究