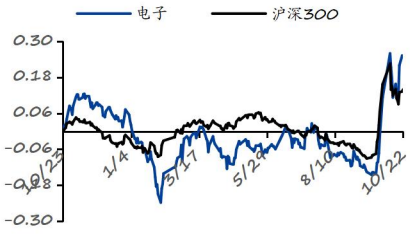
电子

TPU：为更专用的 AI 计算而生

强于大市（维持评级）

一年内行业相对大盘走势



团队成员

分析师： **陈海进(S0210524060003)**
chj30590@hfzq.com.cn
分析师： **徐巡(S0210524060004)**
xx30511@hfzq.com.cn
联系人： **李雅文(S0210124040076)**
lyw30508@hfzq.com.cn

相关报告

1、如何测算文本大模型 AI 训练端算力需求？——2024.6.3
2、从训练到推理：算力芯片需求的华丽转身——2024.8.24
3、Scale Out & Scale Up 兼论，以太网及超节点下数据中心硬件的投资机遇——2024.7.4

投资要点：

- **TPU：为更专用的 AI 计算而生，引领 AI 芯片时代**

作为一种 AI 芯片，TPU 是专用集成电路（ASIC）的代表。主流 AI 芯片架构包括 GPGPU、ASIC 和 FPGA，我们一般认为 GPGPU 为改善 CPU 效率而生，而 TPU 可以进一步改善 GPGPU 未优化完全的部分，三者是从通用到专用不断演进的过程。我们通过梳理发现，TPU 在性能功耗比、集群算力利用率上相较于 GPGPU 有较大优势。主要由于：（1）芯片层面：TPU 专为矩阵乘法而设计，脉动阵列、低精度等设定均适用于 AI 算法，能够处理大量数据以及复杂的神经网络；（2）集群层面：谷歌自研光学芯片 Palomar，构建集群互连优势；TPU 与 TensorFlow 良好适配，软件与硬件相得益彰，能够发挥出 1+1＞2 的效果。站在当下回望，我们发现，TPU 具备的优势其实最终都形成了 AI 芯片共同的趋势，在优化方向上大同小异，而谷歌的强大在于“前瞻”。

- **谷歌 TPU：量级仅次于 NVIDIA GPU，自建租赁模式吸引头部客户**

尽管谷歌没有对外出售自研的 TPU，但随着 TPU v4（2021 年推出）和大型语言模型的出现，谷歌芯片业务的规模显着增加，23 年 TPU 已经突破了 200 万颗量级。根据 Capvision，谷歌 TPU 70%-80%的算力用于内部业务场景使用，剩余 20%-30%以租赁方式供外使用。据集微网，目前全球已经有多家科技公司使用谷歌的 TPU 芯片。超过 60%获得融资的生成式 AI 初创公司和近 90%生成式 AI 独角兽都在使用谷歌 Cloud 的 AI 基础设施和 Cloud TPU 服务，如 Anthropic、Midjourney、Salesforce、Hugging Face 和 AssemblyAI。苹果自 TPUv3 时代便开始使用 Google TPU+GPU 算力，24 年 7 月，苹果公布其使用了 2048 片 TPUv5p 芯片来训练拥有 27.3 亿参数的设备端模型 AFM-on-device，以及 8192 片 TPUv4 芯片来训练大型服务器端模型 AFM-server。

- **国产 TPU：中昊芯英初露锋芒，国产 TPU 登上舞台**

除谷歌外，国产 TPU 厂商中昊芯英逐渐崭露头角。从产品上看，公司首款 TPU 芯片刹那已于 23 年底量产，为国内 AI 产业提供自主可控方案。从业绩上看，公司是国内 AI 芯片唯二盈利的企业（另一个为华为海思），自我造血能力强。据公司 CEO 表示，公司 23 年实现了 4.85 亿的营收和 8000 万的净利润，已通过商业化的阶段性成功逐步实现了自我造血的能力。我们分析其成长路径，自身实力与需求因素共同驱动，不可或缺：一方面，公司 CEO 曾在 Google 作为芯片研发核心团队唯一的华人研发 leader 深度参与 TPU 2/3/4 的设计与研发，履历丰富，团队阵容豪华。另一方面，公司已获青海“丝绸云谷“绿色算力项目订单（首批订单超 9 亿元），并将联手深圳联通共建广东首个国产 TPU 智算中心，大订单支撑公司早期成长。展望未来，我们认为国产智算中心会是一个庞大的算力市场，国产 TPU 有望大展宏图。

- **建议关注**

  - 海外 TPU 龙头：谷歌
  - 国产 TPU 厂商：中昊芯英（未上市）、艾布鲁

- **风险提示**

AI 需求不及预期的风险；TPU 技术升级不及预期的风险；市场竞争加剧的风险。

正文目录

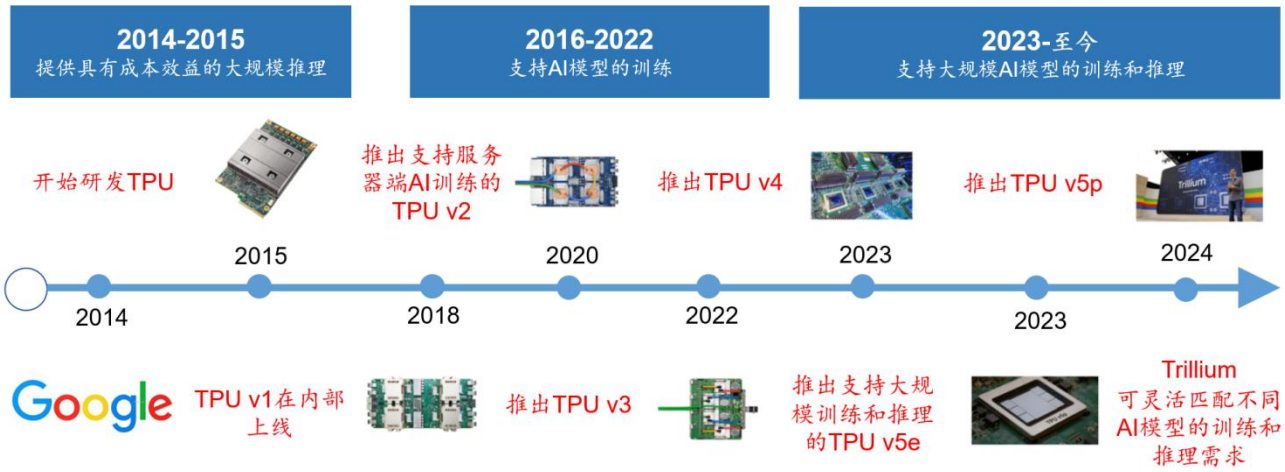
图表目录

1 十年磨一剑，TPU 引领 AI 芯片时代

1.1 TPU 如何发展而来？

简而言之，为更专用的 AI 计算而来。2013 年，Google AI 负责人发现，如果有 1 亿安卓用户每天使用手机语音转文字服务 3 分钟，消耗的算力就已是谷歌所有数据中心总算力的两倍。而传统的通用 CPU 以及专攻图形加速、视频渲染等复杂任务 GPU 无法满足深度学习工作负载的巨大需求，同时存在效率较低、专用运算有限等问题。于是，为探索出更具成本效益、节能的机器学习解决方案，谷歌毅然决定自行研发机器学习专用的处理器芯片，并于 2015 年宣布第一代 TPU 芯片（TPU v1）在内部上线，随后开启了长达 10 年的 TPU 更新迭代。

图表 1：谷歌 TPU 发展史



来源：Google，华福证券研究所

作为一种 AI 芯片，TPU 是专用集成电路（ASIC）的代表。主流 AI 芯片架构包括 GPGPU、ASIC 和 FPGA。GPGPU 通用性强，生态完善，GPGPU 的主要供应商英伟达是 AI 市场的绝对龙头，但 GPGPU 存在着成本高等问题；ASIC 虽然算力强大，功耗小，但相较于 GPGPU 在通用计算上稍有欠缺；FPGA 更具灵活性，也具有足够的算力，但相对开发周期长，复杂算法开发难度大，成本昂贵。TPU 专为单一特定目的而设计：用以运行构建 AI 模型所需的独特矩阵和基于矢量的数学运算。**其架构专为矩阵乘法而设计，**这使它们能够处理大量数据以及复杂的神经网络。需要说明的是，我们也看到相关研究将 TPU 归类为 DSA（专用领域架构处理器），因为 ASIC 是加速某一项功能，而 DSA 是加速某一类功能。但总体上 ASIC 和 DSA 的特征较为相仿，本文不作进一步区分。

图表 2：主流 AI 芯片架构及主要厂商

来源：中兴文档，凡亿企业培训，半导体产业纵横，满天芯，元宇宙投融邦，半导体行业观察，中昊芯英科技，芯榜，与非网 eefocus，华福证券研究所

1.2 TPU 优势何在？

1.2.1 芯片层面：能效王者，架构设计之美淋漓尽致

六代版本更新，与 GPGPU 平分秋色。我们将历代 TPU 以及同时代的 GPGPU 进行梳理。首先，我们观察到同代 TPU 与 GPGPU 大多数处于同代或相近制程。第四代 TPU 已采用 7nm 制程，据 The Next Platform 推测第五代/第六代 TPU 分别采用 5nm/4nm 制程，而英伟达 Ampere/Hopper/Blackwell 架构分别采用 7nm/4nm/4nm 制程。**在算力上，谷歌目前暂时落后一代。**2024 年谷歌发布第六代 TPU Trillium，实现最大算力 926TFLOPS（BF16）/1852TFLOPS（INT8），相较于第五代 TPU v5e 和 v5p 实现了飞跃式上升，比肩英伟达 2023 年发布的 H100，对应算力为 989TFLOPS（FP16）/1978TFLOPS（INT8 or FP8）。**但在性能功耗比上，我们认为谷歌优势显著。**谷歌并未披露最新产品的功耗指标，我们从前代产品可以窥见一二——2021 年发布的第四代 TPU v4 性能功耗比为 0.89-1.31TOPS/W，而英伟达同代产品 A100（2020 年发布）的性能功耗比为 1.56TOPS/W。

图表 3：TPU 和 GPGPU 对比图

Google TPU								
芯片	TPU v1	TPU v2	TPU v3	TPU v4i	TPU v4	TPU v5e	TPU v5p	Trillium
发布时间	2015	2017	2018	2020	2021	2023	2023	2024
制程	28	16	16	7	7	5	5	4
BF16 TFLOPS	/	46	123	137.5	69	197	459	926
INT8 TFLOPS	92	/	/	138	275	394	918	1852
显存类型	DDR3	HBM	/	/	HBM2	HBM2	HBM3	HBM3或HBM3e
内存（GB）	8	16	32	8	32	16	95	32
NVIDIA GPU								
芯片		V100		A100		H100	H200	GB200
发布时间		2017		2020		2022	2023	2024
制程		12		7		4	4	4
FP64 TFLOPS		7.8		9.7		33.5	34	/
FP32 TFLOPS		15.7		19.5		67	67	/
FP16 Tensor TFLOPS		125		623.8		989	1979	5000
INT8/FP8 Tensor TFLOPS		/		/		1978	3958	10000
显存类型		HBM2		HBM2e		HBM3	HBM3e	HBM3e
内存（GB）		16		80		80	141	384

来源：The Next Platform，NVIDIA 官网，芯东西，量子位，新智元，AI 时代前沿，IT 之家，AI 科技评论，华福证券研究所
注：第五代和第六代 TPU 制程为 The Next Platform 推测

对此现象，我们可以从逻辑芯片架构的角度来进行解释——通用处理器 CPU 和 GPGPU 因架构设计而在 AI 计算上存在低效问题。我们一般认为 GPGPU 为改善 CPU 效率而生，而 TPU 可以进一步改善 GPGPU 未优化完全的部分，三者是从通用到专用不断演进的过程。据新智元，CPU 使用了非常大量的片上存储来做缓存（Cache），将程序经常访问的数据放在片上，这样就不必访问内存了，从而实现“内存访问近乎零延迟”，相比之下负责运算的算术逻辑单（ALU）只占据了一小部分，这就是 CPU 进行大规模并行数据运算时效率低的原因之一。GPU 里面有数千个小核心，每个都可以看成是个小 CPU，它可同时运行最多数十万个小程序。虽然 GPU 单核的处理能力弱于 CPU，但是数量庞大，ALU 占比大，非常适合高强度并行计算。但实际上大多数程序会因为等待访存而卡住，且管理和组织大量程序会付出巨大的硅片面积代价和内存带宽的代价，这个是 GPU 低效的根源。

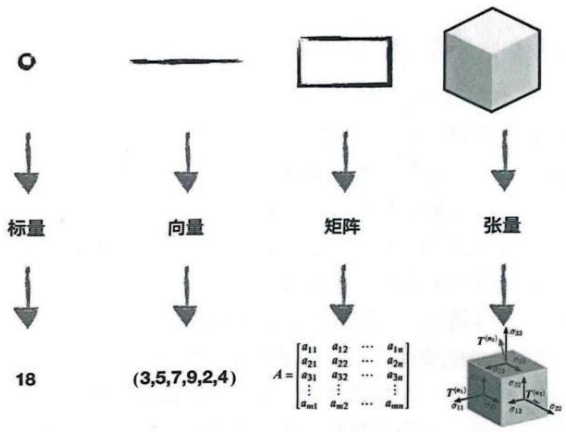
图表 4：通用处理器 CPU 和 GPU 的逻辑架构

来源：EDA365 电子论坛，华福证券研究所

TPU 优势#1 脉动阵列为基，张量计算横空出世——增大吞吐量，节省时间

TPU 本义为张量处理器，这其中的“张量”是在数学和物理领域常见的概念。从定义上讲，张量是一个多维数组，可以具有任意维度，张量的元素可以是标量、向量或更高维度的张量。一个数值可以看作一个标量（零维张量），一个一维数组可看作一个向量（一维张量），一个二维数组是一个矩阵（二维张量）。在深度学习和神经网络中，**我们经常用高维的张量来表示图像、音频、文本等数据。**通过在神经网络的各个层级之间进行张量的传递和计算，神经网络能学习和处理复杂的输入数据，执行特征提取、分类、回归等任务。

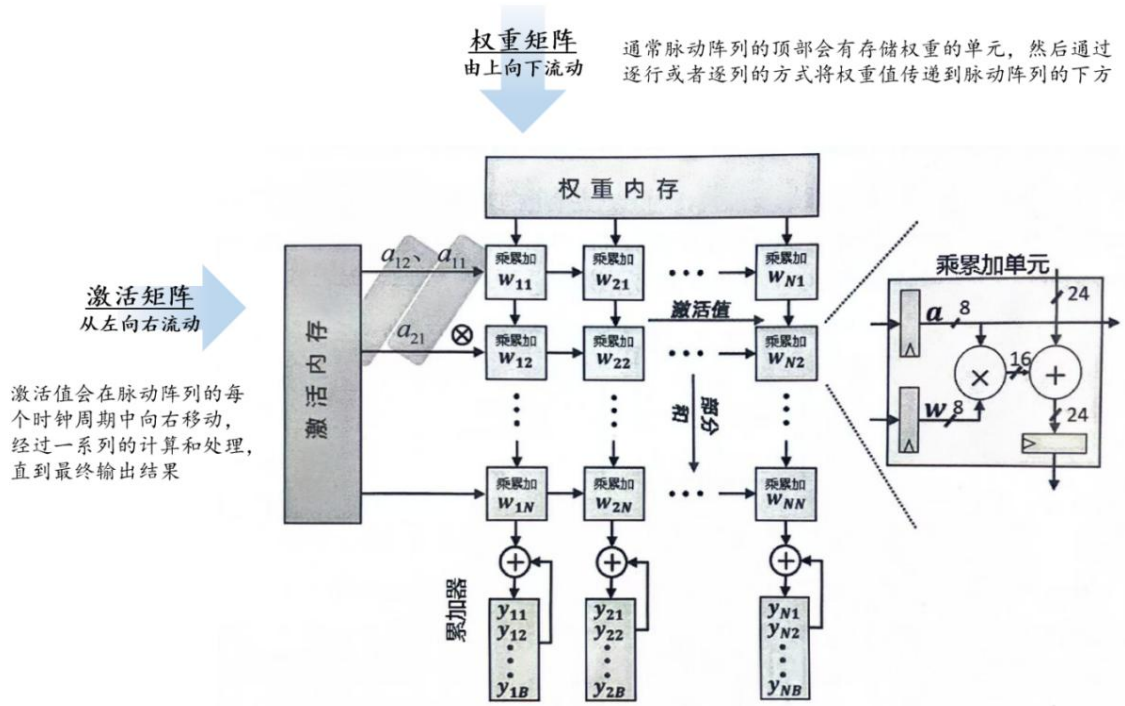
图表 5：张量的图形化表达



来源：濮元恺《算力芯片——高性能 CPU/GPU/NPU 微架构分析》，华福证券研究所

TPU 的核心是 MXU（矩阵乘法单元），MXU 以脉动阵列为架构，使 TPU 能够以很高的吞吐量执行矩阵乘法和累加。脉动阵列作为 TPU 的底层技术，是一种适用于进行大量的并行计算（尤其是矩阵乘法，也是深度学习中最常见的操作）的计算硬件结构。脉动阵列的名字来源于它的工作方式，即数据在阵列中“脉动”式地流动，就像心脏在血管中泵血一样。通过这种方式，脉动阵列可以高效地执行矩阵计算操作，因为数据的流动方向符合计算规则和数据依赖关系。这种并行的数据流动方式可以充分利用硬件结构的并行性，加速矩阵计算过程。**脉动阵列也必然存在局限性，**比如它的计算模式相对固定，不适合执行有大量控制流的计算。不过，在深度学习中，大部分的计算都是数据流式的，且执行并行的矩阵计算，因此这个局限性的影响并不大。

图表 6：脉动阵列模型与一个乘累加单元



来源：濮元恺《算力芯片——高性能 CPU/GPU/NPU 微架构分析》，华福证券研究所

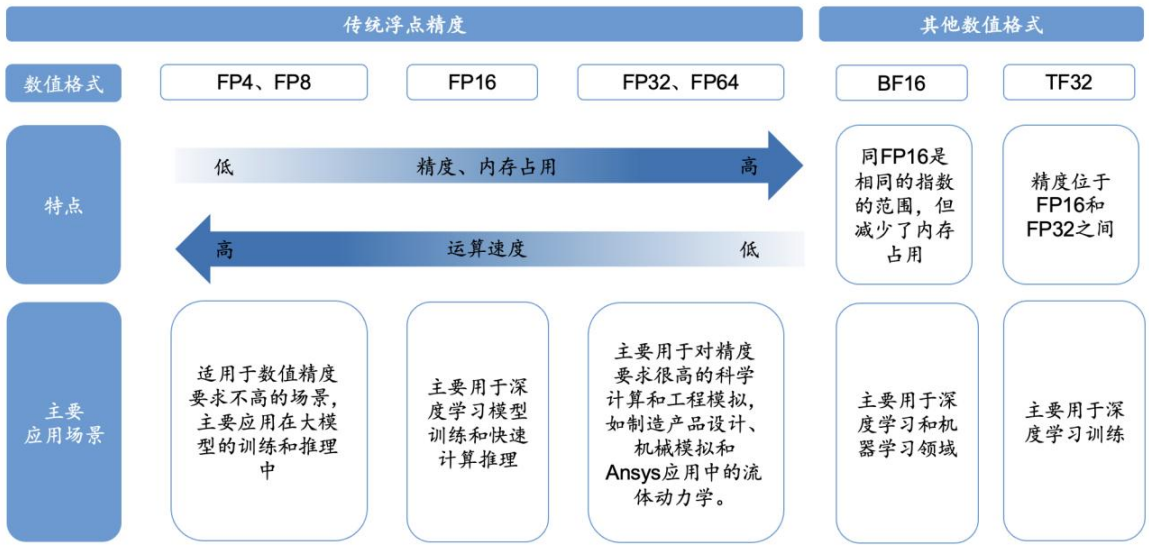
TPU 优势#2 直击 AI 应用，聚焦低精度计算——节省芯片面积

回看本文开篇对谷歌发展历程的复盘，大致可以将其划分为两个时代——以 ChatGPT 为分水岭的 AI 初探索阶段和 AI 大爆发时代。TPU 自发明以来一直以低精度著称，从 AI 初探索阶段迈入 AI 大爆发时代，TPU 也经历了“拨云见日”的过程。

AI 初探索阶段：初代 TPU 入局 AI 推理，TPU v2 入局 AI 训练——低精度计算即可满足 AI 计算需求，特定方向优化出奇效。神经网络的两个主要阶段是训练（Training 或者学习 Learning）和推理（inference 或者预测 Prediction）。实际上，初代 TPU 推出的同一时期，训练几乎都是基于浮点运行，这也是 GPU 流行的原因之一。事实上，推理过程使用 INT8 也基本够用。**INT8 运算相较于浮点数而言，对整个芯片的能耗、面积都有较大程度上的节省，主要包括以下两方面：**（1）乘法运算：INT8 乘法比 IEEE 754 标准下 FP16 乘法降低 6 倍的能耗，占用的硅片面积也少 6 倍；（2）加法运算：整数加法的收益是 13 倍的能耗与 38 倍的面积。从这一角度出发，谷歌 TPU 设计顺势采用低精度计算模式，TPU v1 仅支持 INT8 精度，而同时代英伟达的 AI 推理芯片 K80（2014 年推出）最低需要支持 FP32 精度。实际效果显示，与 GPU 相比，TPU 的控制逻辑单元更小，更容易设计，面积只占整体芯片面积的 2%，给片上存储器和矩阵计算单元留下了更大的空间。后来从 TPU v2 开始，谷歌引入了自创的浮点精度 BF16，虽与 FP16 保持相同位数（在浮点精度的位数上与英伟达同时代产品 V100 保持了一致），但能够减少内存占用，也对 AI 硬件的发展产生深远影响。

在这一阶段，AI 应用的方向还不够清晰，AI 硬件的发展路径也并不明朗。TPU 是谷歌基于自身业务以及对 AI 的理解而做出的选择。对于低精度的聚焦，既是在 AI 计算上的优势，同样也是在其他计算领域的劣势。这也是 ASIC 本身的专用化特征所造就的。站在当下回望，**我们发现，TPU 具备的优势其实最终都形成了 AI 芯片共同的趋势，在优化方向上大同小异，而谷歌的强大在于“前瞻”。**

图表 7：浮点精度的特点与应用场景



来源：广州恒联，联泰集群 LTHPC，华福证券研究所

AI 大爆发时代：AI 应用大势所趋，低精度运算成为大规模 AI 训练&推理的标签特征——TPU v5 支持大规模训练推理水到渠成。随着 AI 应用来到“ChatGPT”时刻，大语言模型达到数万亿参数，大规模 AI 计算时代已经到来。AI、高性能计算和数据分析变得日益复杂，AI 模型厂商有时愿意牺牲精度值来获取大模型训练的计算能力，也就是用高运算速度、低存储需求来加速计算过程。通过梳理我们发现，不止 TPU，GPGPU 也在向着低精度趋势发展。

图表 8：部分人工智能芯片支持的数值格式

产品	昇腾910	A100	MI250X	H100	MI300A	Gaudi 2	GB200	Gaudi 3	B200	MI350	TPU v1	TPU v2	TPU v3	TPU v4	TPU v5e	TPU v5p	Trillium
厂商	华为	NVIDIA	AMD	NVIDIA	AMD	Intel	NVIDIA	Intel	NVIDIA	AMD	Google	Google	Google	Google	Google	Google	Google
发布时间	2019年8月	2020年5月	2021年11月	2022年3月	2023年6月	2023年7月	2024年3月	2024年4月	2024年8月	计划于2025年发布	2016年5月	2017年5月	2018年5月	2021年5月	2023年8月	2023年12月	2024年5月
FP4							✓		✓	✓							
FP6							✓		✓	✓							
FP8				✓	✓	✓	✓	✓	✓	✓							
FP16	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓							
FP32		✓	✓	✓		✓	✓	✓	✓								
FP64		✓	✓	✓			✓		✓								
BF16		✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
TF32		✓		✓	✓	✓	✓	✓	✓	✓							
INT8	✓	✓	✓	✓	✓		✓		✓	✓	✓			✓	✓	✓	✓

来源：各公司官网，The Next Platform，IT 之家，机器之心，华福证券研究所

1.2.2 集群层面：算力利用率是最好的证明

在 AI 大模型预训练方面，谷歌 TPU 的算力利用率表现明显领先于英伟达。据量子位，算力利用率（MFU）是实际吞吐量与理论最大吞吐量之比。训练大语言模型并非简单的并行任务，需要在多个 GPU 之间分布模型，并且这些 GPU 需要频繁通信才能共同推进训练进程。通信之外，操作符优化、数据预处理和 GPU 内存消耗等因素，都对算力利用率（MFU）这个衡量训练效率的指标有影响。GPT-3 到 GPT-4 明显看到算力利用率由 21.3%提升至 34%（32-36%区间，本文取中值粗略计算），趋势上较为明确。横向对比发现，相较于 OpenAI 的 GPT 系列，谷歌利用 TPU 训练的 Gropher 和 PaLM 明显在算力利用率上更胜一筹，我们认为谷歌自研 TPU 在自有大模型训练上展现出独特的优势。

图表 9：算力利用率

	单位	GPT3	GPT4	Gopher	PaLM
发布机构		OpenAI	OpenAI	DeepMind	Google
发布时间		2020-05	2023-03	2021-12	2022-04
算力供给厂商		英伟达	英伟达	谷歌	谷歌
GPU产品		V100	A100	TPU v3	TPU v4
GPU数量	片	10000	25000	4096	6144
理论峰值FP16 TC	TFlops	125	312		
算力利用率		21%	34%	33%	46%

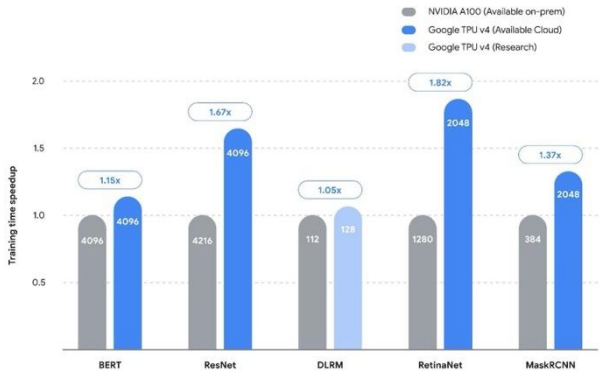
来源：Aakanksha Chowdhery 等《PaLM : Scaling Language Modeling with Pathways》，英伟达，谷歌研究院，腾讯科技，机器之心，机器之心 Pro，新智元，中关村在线，河北省科学技术厅，华福证券研究所

注：GPT4 算力利用率在 32-36%区间，本文取中值粗略计算

从训练效果上看，谷歌 TPU 也有不俗表现。22 年发布的第四代 TPU 在许多 MLPerf 基准测试（最显著的是深度学习和卷积网络）上的表现优于英伟达。在 MLPerf 五项基准测试中，TPUv4 性能比 A100 高出 40%。同时，在各 AI 厂商关注的训练成

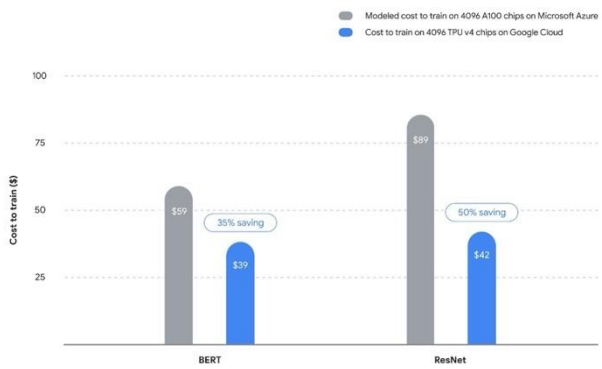
本上，谷歌的 TPU v4 相较于 A100 表现更优异。

图表 10：TPU v4 和 A100 在各种模型上的训练效果



来源：The Next Platform，华福证券研究所

图表 11：TPU v4 和 A100 在训练上的成本比较

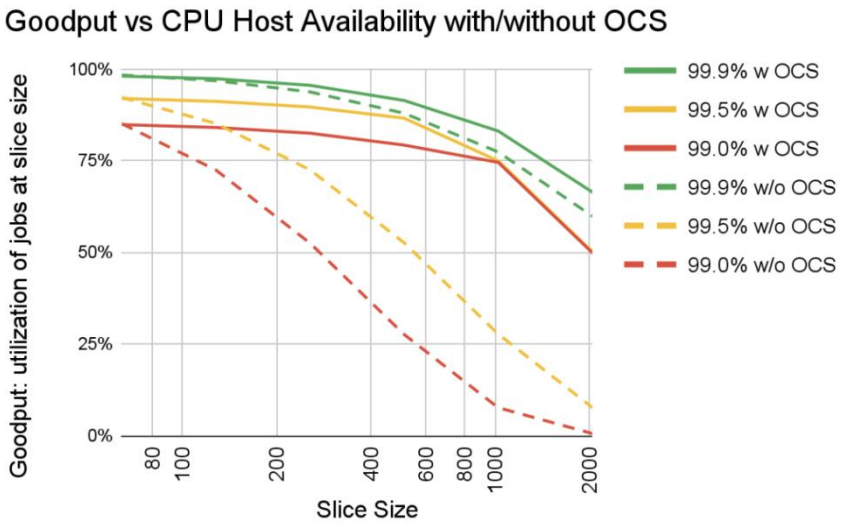


来源：The Next Platform，华福证券研究所

对此现象，我们可以从硬件/软件多个角度来进行解释——

#硬件优势 谷歌自研光学芯片 Palomar，从集群互连角度构建优势。谷歌设计 TPU 的目的就是构建自己的超级计算机，如何高速度、低延迟地把尽可能多的 TPU 芯片连接起来是一个不可避免的问题。谷歌又一次把握前瞻方向，在常规的互连拓扑结构中罕见地自研了光学芯片 Palomar（谷歌 TPU v4 设计的其中一个重点），使用该芯片实现了全球首个数据中心级的可重配置 OCS。在 Palomar 芯片加入后，立方体结构节点之间的互连并非一成不变的，而是可以现场重配置，这样做的最大好处是可以根据具体的机器学习模型来改变拓扑，以及改善超级计算机的可靠性。如下图所示，在使用可重配置光互连（以及光路开关时），系统有效吞吐量和利用率大幅提升。

图表 12：谷歌自研光学芯片 Palomar 的性能



来源：Norman P. Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, Cliff Young, Xiang Zhou, Zongwei Zhou, and David Patterson Google e, Mountain View, CA《TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with》，华福证券研究所

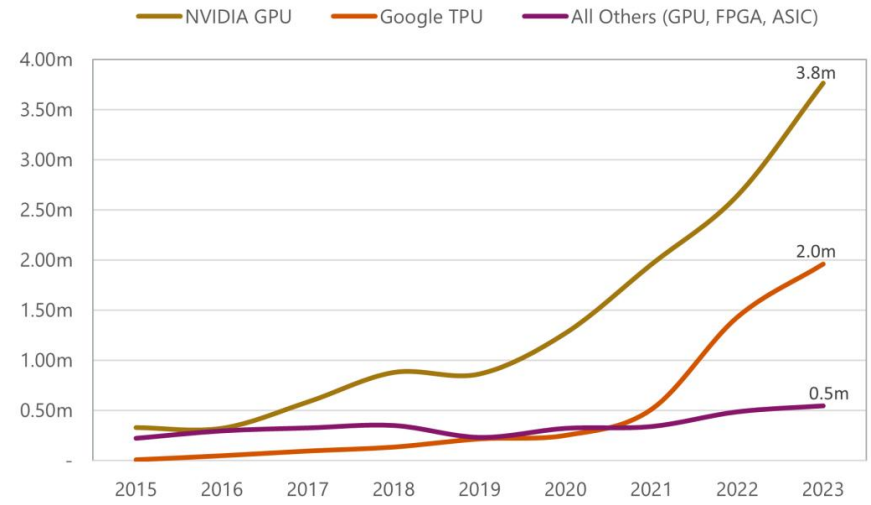
注：横轴为分片大小，纵轴为作业在分片大小时的有效吞吐量和利用率

#软件优势 TPU 专为 TensorFlow 打造，软件与硬件相辅相成。TensorFlow 是 Google 的一个开源机器学习软件库。TPU 是根据 TensorFlow 设计的，从而能够降低运算精度，在相同时间内处理更复杂、更强大的机器学习模型并将其更快投入使用。初代 TPU 的设定，只能在 TensorFlow 中执行推理，但它的性能非常好。我们认为，TPU 与 TensorFlow 的良好适配，能够发挥出 1+1＞2 的效果。深度学习计算中的芯片部署都不是零和博弈。现实世界的深度学习网络需要系统的 GPU 与其他 GPU 或诸如 Google TPU 之类的 ASIC 通信。GPU 是理想的工作环境，具有深度学习所需的灵活性。但是，当完全专用于某个软件库或平台时，则 ASIC 是最理想的，谷歌 TPU 与 TensorFlow 就是最好的例子。

2 谷歌视角：如何理解 TPU 的生态位？

年出货 200 万颗，量级仅次于 NVIDIA GPU。自 2015 年谷歌推出自研的 TPU 以来，尽管谷歌没有对外出售自研的 TPU，但随着谷歌 TPU 的不断发展，其出货量随着每一代新 TPU 的推出而加速增长。随着 TPU v4（2021 年推出）和大型语言模型的出现，谷歌芯片业务的规模显着增加，23 年 TPU 已经突破了 200 万颗量级。

图表 13：全球数据中心加速器年出货量

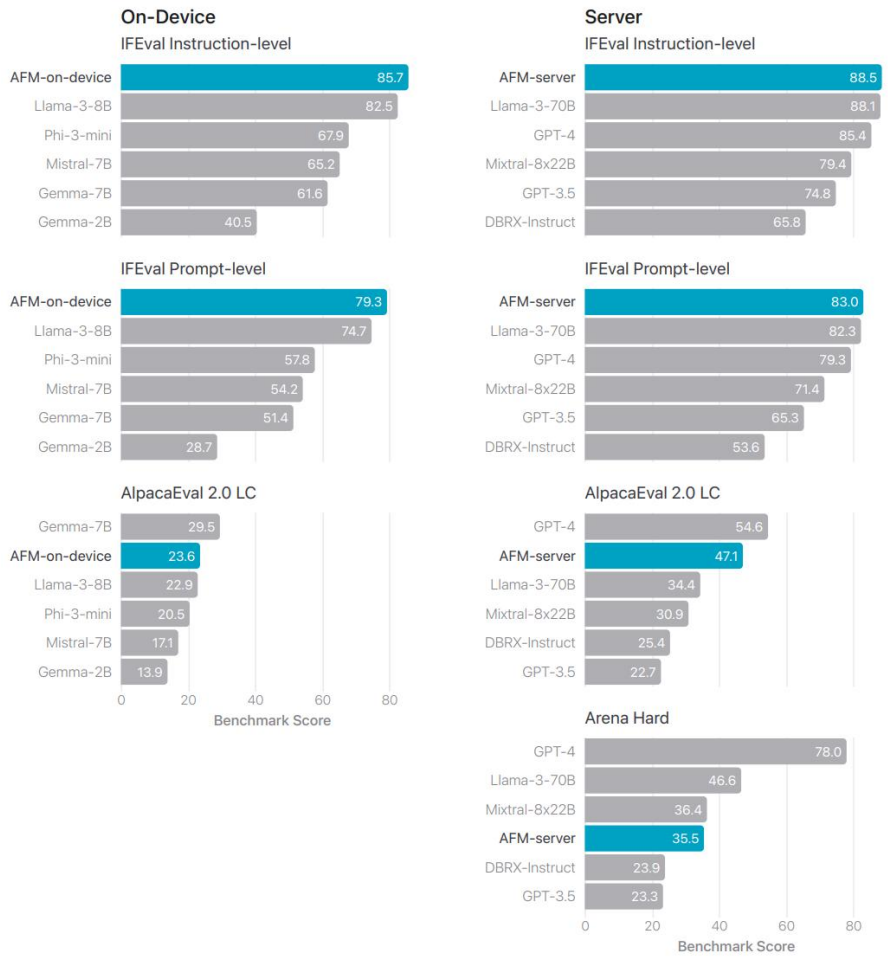


来源：TechInsights，华福证券研究所

不对外售卖，以租赁方式作为主要盈利来源。根据 Capvision，Google AI 算力一直采用自建租赁模式。自家 TPU 平均每 1-1.5 年更新一代，其中 70%-80%的算力用于内部业务场景（搜索，广告，视频，Gemini 等）使用，剩余 20%-30%以租赁方式供外使用。据集微网，目前全球已经有多家科技公司使用谷歌的 TPU 芯片。超过 60%获得融资的生成式 AI 初创公司和近 90%生成式 AI 独角兽都在使用谷歌 Cloud 的 AI 基础设施和 Cloud TPU 服务，并广泛应用于社会经济各个领域。例如 Anthropic、Midjourney、Salesforce、Hugging Face 和 AssemblyAI 等知名 AI 创企在大量使用 Cloud TPU。24 年 7 月，苹果公布其使用了 2048 片 TPUv5p 芯片来训练拥有 27.3 亿参数的

设备端模型 AFM-on-device，以及 8192 片 TPUv4 芯片来训练其为私有云计算环境量身定制的大型服务器端模型 AFM-server。Apple 整体战略部署上早年就意识到 NVIDIA CUDA 闭源模式弊端不利于自身生态的长远发展，自 TPUv3 时代便开始使用 Google TPU+GPU 算力，并借助 JAX 开源平台与 CUDA 底层逻辑相似性，实现快速部署迁移训练（Google 开源的 JAX 以其与 CUDA 相似的开发框架逻辑，降低用户学习与迁移成本，实现市场突破）。另加之 2023 年全球市场 N 卡排队，进一步促进双方合作。

图表 14：AFM 模型和其他模型性能对比



来源：Apple《Apple Intelligence Foundation Language Models》，华福证券研究所

3 TPU 商业模式何解？

3.1 为什么谷歌 TPU 能够成功？

首先，自产自销，算力与算法紧密联结。若将 TPU 理解为 ASIC 芯片的一种，ASIC 作为依产品需求不同而定制化的特殊规格集成电路，具有高性能、低功耗优势，但它们只能执行特定算法。ASIC 生产成本高，因此当出货量较小时，采用 ASIC 并不经济。而当需求开始增加，芯片出货量增加，可通过芯片代工厂批量生产来降低成本。谷歌能够预估和规划自己在哪个阶段需要多少计算资源，也清晰地知道通用

GPU 上的哪些功能单元是用不到的，这也是为什么谷歌不选择使用现有的 GPU 芯片，而是自行研发生产 TPU 的重要原因。

其次，自身能力过硬，对 AI 理解深入且前瞻。业界普遍认为没有厂商比谷歌更懂 AI 用户的需求，所以谷歌根据自己的需求定制化开发的 TPU 芯片一经发布就引发了大量讨论。站在此时去回看，也会发现谷歌引领了 AI 芯片的发展方向。

TPU 浪潮初现，更多 AI 科技龙头开始探索 TPU 或类 TPU。除谷歌外，全球越来越多的顶尖科技公司开始研发或使用 TPU 或类 TPU 架构的 AI 专用芯片。科技巨头们在尖端 AI 训练方面开始寻求更多元化的解决方案的趋势。

图表 15：顶尖公司对 TPU 或类 TPU 的探索

公司	AWS	Meta	特斯拉	微软	AWS	Meta	英特尔	OpenAI
时间	2020年12月	2023年5月	2023年7月	2023年11月	2023年11月	2024年4月	2024年4月	2024年6月
事件	发布专用于训练机器学习的芯片Trainium	官宣自研的采用RISC-V开源架构的AI算力芯片MTIA	马斯克暗示特斯拉正在开发与传统GPU不同架构的芯片	推出专为Azure云服务和AI工作负载设计的ASIC芯片Maia 100	发布为生成式AI和机器学习训练设计的云端AI芯片Trainium 2	推出第二代自研AI训练和推理芯片MTIA v2	推出专供深度学习神经网络推理的类TPU芯片Gaudi 3	SemiAnalys报道其正从谷歌TPU团队招募研发人才，并开始自研AI芯片

来源：中昊芯英科技，智东西，新智元，集微网，华福证券研究所

3.2 国产 TPU 厂商中昊芯英崭露头角

在 AI 芯片领域，TPU 厂商并不多见，除谷歌外，国产 TPU 厂商中昊芯英逐渐崭露头角。从产品上看，中昊芯英首款 TPU 已量产，为国内 AI 产业提供自主可控方案。中昊芯英历时近五年研发的刹那 TPU AI 芯片已于 23 年底实现量产，拥有完全自主可控的 IP 核、全自研指令集与计算平台。在处理大规模 AI 模型计算任务时，相较于英伟达 2020 年推出的 A100，刹那的计算性能超越其近 1.5 倍，在完成相同 AI 大模型计算任务量时的能耗降低 30%，单位算力成本仅为其 42%。刹那以独特的高达 1024 片芯片高速片间互联的能力构建大规模智算集群泰则，系统集群性能远超传统 GPU 数十倍，可支撑超千亿参数 AIGC 大模型计算需求。该成就不仅打破了国外企业在高端 AI 芯片领域的垄断地位，更为国内 AI 产业的发展提供了自主可控的解决方案，解决了卡脖子难题。

图表 16：产品性能比对图

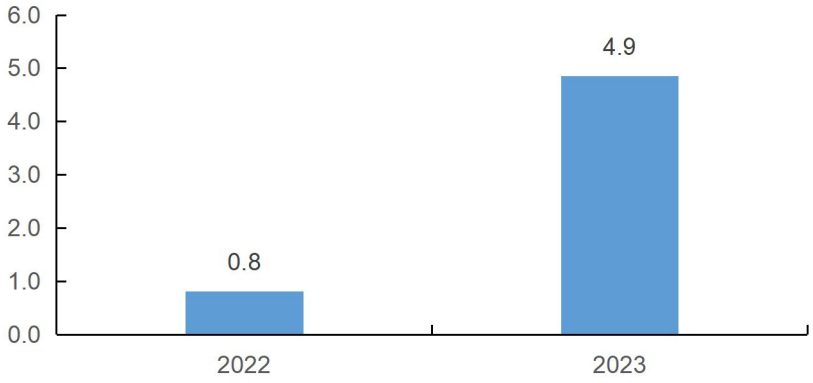
芯片	刹那®	TPU v4	TPU v5p	A100 SXM
厂商	中昊芯英	Google	Google	NVIDIA
发布时间	2023	2021	2023	2020
算力	超越A100近1.5倍	275TFLOPS（BF16 or INT8）	459TFLOPS（BF16）/918TOPS（INT8）	312TFLOPS（BF16 or FP16)/624TOPS（INT8）
最多互联数量	1024	4096	8960	/

来源：中昊芯英科技，芯智讯，英伟达官网，华福证券研究所

从业绩上看，中昊芯英是国内 AI 芯片唯二盈利的企业（另一个为华为海思），自我造血能力强。据中昊芯英 CEO 表示，公司 23 年实现了 4.85 亿的营收和 8000 万的净利润，已通过商业化的阶段性成功逐步实现了自我造血的能力，不再单纯依靠融资存活。按照中昊芯英股东科德教育与其签订的对赌协议，中昊芯英 2023-2024 年两年合计销售总收入不得低于 7.6 亿元（其中 2023 年销售总收入不得低于 2.08 亿

元），此外，还要求公司在 2026 年 12 月 31 日前完成 IPO 或被收购。

图表 17：中昊芯英营收（单位：亿元）



来源：科德教育公告，华福证券研究所

我们分析其成长路径，自身实力与需求因素共同驱动，不可或缺：

一方面，中昊芯英是国产 TPU 芯片独角兽，团队阵容豪华。中昊芯英成立于 2018 年，是国内唯一掌握 TPU 架构训推一体 AI 芯片核心技术的公司。公司核心团队由来自谷歌、微软、甲骨文、三星、英伟达、亚马逊等顶尖科技公司的 AI 软硬件设计专家组成，团队成员拥有从 28nm 到 7nm 芯片设计、优化、流片生产、客户交付的完整方法论与全流程经验。公司创始人及 CEO 杨龚轶凡毕业于斯坦福大学，随后在美国硅谷从事了 10 余年高端芯片领域的研发工作，在 Oracle 参与、主导了 12 款顶级高性能 CPU 的设计与产出，在 Google 作为芯片研发核心团队唯一的华人研发 leader 深度参与 TPU 2/3/4 的设计与研发，产业生涯成功流片 10 余次。

另一方面，大订单支撑中昊芯英早期成长。中昊芯英已获青海“丝绸云谷“绿色算力项目订单（首批订单超 9 亿元），并将联手深圳联通共建广东首个国产 TPU 智算中心。**（1）青海“丝绸云谷”：**“丝绸云谷”低碳算力产业园项目以打造西北数据云谷为终极目标，是青海首个万卡集群项目，**总投资约 230 亿元**，分两期建设（一期建设时间为 2023 年至 2025 年，二期建设时间为 2026 年至 2028 年）。**园区建成后将容纳约 20 万台高性能 AI 服务器运行，**有望成为国内最大规模零碳数据中心余热回收利用一体化项目，也是国内首个完全定位于“大算力+大模型”形态的大型 AI 计算中心。该项目计划分批采购中昊芯英自研 AI 训练服务器及计算集群产品以搭建 AI 计算底座，首批规划订单需求超 9 亿元。**（2）广东国产 TPU 智算中心：**中昊芯英和深圳联通于 24 年 9 月宣布携手合作，共同建设广东首个采用国产 TPU 技术的智算中心。该项目一期由 32 个算力节点通过高效互联构建而成，**整体算力不低于 50P，后期将扩容至千卡规模，**形成训推一体化的枢纽，成为中国联通在深圳的核心智算高地的重要组成部分。

展望未来，我们认为国产智算中心会是一个庞大的算力市场，也是一个国产 AI 芯片公司可以大展宏图的地方。从客户角度分析，我们认为智算中心客户不像互联网大厂具有自研 AI 芯片的能力，一定程度上也是对第三方芯片公司非常友好的市场。

而全国各地的智算中心需求加在一起也是一个足够庞大的算力市场，完全有机会分摊掉 TPU/ASIC 的研发成本，也具备商业合理性。综合来看，是一种较为可行的商业化落地思路。

图表 18：各省算力规划

省份	文件名	发布时间	计划建成算力
福建	《福建省新型基础设施建设三年行动计划(2023—2025年)》	2023年7月	2024年6.5EFLOPS，2025年8EFLOPS
湖北	《湖北省加快发展算力与大数据产业 三年行动方案（2023-2025年）》	2023年8月	2025年8EFLOS
重庆	《重庆市算力网络发展“算力山城 强算赋能”行动计划（2023-2025年）》	2023年12月	2024年8EFLOPS，2025年10EFLOPS
山西	《山西省算力基础设施高质量发展实施方案 》	2024年1月	2025年9EFLOPS
青海	《青海省绿色算力基地建设方案》	2024年2月	2025年2.06EFLOPS
上海	《上海市智能算力基础设施高质量发展“算力浦江”智算行动实施方案(2024-2025年)	2024年3月	2025年30EFLOPS
浙江	《浙江省智能物联产业集群建设行动方案》	2024年3月	2027年40EFLOPS
河南	《河南省加快制造业“六新”突破实施方案》	2024年3月	2025年2EFLOPS
广东	《广东省算力基础设施高质量发展行动暨“粤算”行动计划（2024-2025年）》	2024年3月	2024年28EFLOPS，2025年38EFLOPS
北京	《北京市算力基础设施建设实施方案（2024-2027年）》	2024年4月	2025年45EFLOPS
江苏	《江苏省算力基础设施发展专项规划》	2024年4月	2025年24EFLOPS，2030年50EFLOPS
陕西	《陕西省加快推动人工智能产业发展实施方案(2024-2026年)》	2024年5月	2026年3EFLOPS
河北	《河北省人民政府办公厅关于进一步优化算力布局推动人工智能产业创新发展的意见》	2024年5月	2025年35EFLOPS
山东	《山东省算力基础设施高质量发展行动方案》	2024年5月	2025年12.5EFLOPS
甘肃	《甘肃算力基础设施高质量发展三年行动计划（2024-2026年）》	2024年5月	2026年30EFLOPS
安徽	《陕西省加快推动人工智能产业发展实施方案(2024-2026年)》	2024年5月	2025年12EFLOPS
西藏	《“算力珠峰”高质量发展行动计划（2024-2026）》	2024年6月	2026年0.1EFLOPS
天津	《天津市算力产业发展实施方案（2024—2026年）》	2024年7月	2026年10EFLOPS
贵州	《贵州省“千兆黔省、万兆筑城”行动计划 (2024-2025年)》	2024年7月	2024年70EFLOPS，2025年200EFLOPS

来源：各省政府官方，上海证券报，中国证券网，浙江省民营经济研究中心，新京报，湖北网络广播电视台，华福证券研究所
注：所列数据为总的算力规模，包含智算和超算算力

4 风险提示

AI 需求不及预期的风险；TPU 技术升级不及预期的风险；市场竞争加剧的风险。

分析师声明

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

一般声明

华福证券有限责任公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告的信息均来源于本公司认为可信的公开资料，该等公开资料的准确性及完整性由其发布者负责，本公司及其研究人员对该等信息不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，之后可能会随情况的变化而调整。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

在任何情况下，本报告所载的信息或所做出的任何建议、意见及推测并不构成所述证券买卖的出价或询价，也不构成对所述金融产品、产品发行或管理人作出任何形式的保证。在任何情况下，本公司仅承诺以勤勉的职业态度，独立、客观地出具本报告以供投资者参考，但不就本报告中的任何内容对任何投资做出任何形式的承诺或担保。投资者应自行决策，自担投资风险。

特别声明

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。投资者请勿将本报告视为投资或其他决定的唯一参考依据。

投资评级声明

类别	评级	评级说明
公司评级	买入	未来 6 个月内，个股相对市场基准指数指数涨幅在 20%以上
	持有	未来 6 个月内，个股相对市场基准指数指数涨幅介于 10%与 20%之间
	中性	未来 6 个月内，个股相对市场基准指数涨幅介于-10%与 10%之间
	回避	未来 6 个月内，个股相对市场基准指数涨幅介于-20%与-10%之间
	卖出	未来 6 个月内，个股相对市场基准指数涨幅在-20%以下
行业评级	强于大市	未来 6 个月内，行业整体回报高于市场基准指数 5%以上
	跟随大市	未来 6 个月内，行业整体回报介于市场基准指数-5%与 5%之间
	弱于大市	未来 6 个月内，行业整体回报低于市场基准指数-5%以下

备注：评级标准为报告发布日后的 6~12 个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中 A 股市场以沪深 300 指数为基准；香港市场以恒生指数为基准，美股市场以标普 500 指数或纳斯达克综合指数为基准（另有说明的除外）

联系方式

华福证券研究所 上海
公司地址：上海市浦东新区浦明路 1436 号陆家嘴滨江中心 MT 座 20 层
邮编：200120
邮箱：hfyjs@hfzq.com.cn