计算机行业2025年7月投资策略

AI ASIC市场规模快速增长，稳定币产业链蓄势待发

优于大市

核心观点

AI ASIC：价格、功耗优势显著，市场规模快速增长。1）**价格优势**：由于GPU芯片强大的通用性、灵活性，其设计及流片成本较高，进而导致GPU的平均单价较高;根据IDC统计数据,2024年GPU平均单价为8001美元,AI Asic平均单价为5236美元，AI ASIC具备价格优势；2）**功耗优势**：由于AI AISC芯片偏定制化设计，专为特定任务（例如AI大模型的训练或推理）优化，因此其在执行特定任务时功率较低。3）**市场规模**：根据IDC披露数据，2024年GPU、AI ASIC芯片市场规模分别为701、148亿美金，预计2030年分别增长至3263、838亿美金，对应24-30年CAGR分别为29.2%、33.5%。从出货量来看，2024年GPU、AI ASIC芯片出货量分别为876、283万颗，预计2030年增长至2982、1431万颗，对应24-30年CAGR分别为22.6%、31.0%，AI ASIC芯片占比稳步提升。分领域来看，ASIC芯片在训练、训练&推理双用AI芯片领域，增速快于GPU。

复盘谷歌TPU发展历程,AI ASIC三大发展趋势逐步明朗。我们对全球AI ASIC龙头谷歌TPU进行复盘，AI ASIC芯片发展呈现三大发展趋势。1）**专用性持续增强，颗粒度更细**：特别是TPU v5分为了TPU v5e和TPU v5p两个版本，其中TPU v5e是训推一体，强调成本效益，而TPU v5p性能强劲，专注于超大基础模型训练，芯片应用场景更细分；2）**更强的算力、HBM和集群能力**：单卡算力持续提升，选用更领先的HBM（应对多模态任务），单POD芯片数量持续提升，集群拓展效率逐步接近线性；3）**能效比持续提升**：以单芯片封装每瓦热设计功耗所提供的峰值FP8 Flops衡量，Ironwood峰值能效是上一代Trillium的2倍，是TPU v2的29.3倍；同时，TPU v3开始配套液冷，液冷等新一代冷却方式逐步应用。

稳定币：香港政策落地，关注板块投资机会。香港于2025年5月21日正式通过《稳定币条例》，解决稳定币行业长期存在的透明度不足、赎回风险等问题，同时为合规机构开辟清晰的入场路径。1）**提升跨境支付效率和普惠性**：稳定币在交易速度、成本以及体制方面显著优于传统跨境支付体系，绕开SWIFT体系的低效性，重构全球金融包容性格局；2）**赋能RWA资产链上化与全球流通**：稳定币通过价值锚定、效率革命、合规护航、流动性激活，破解RWA项目的价格波动、跨境摩擦、信任缺失与门槛高等痛点。

投资建议:看好AI ASIC及稳定币。谷歌、亚马逊、Meta等公司纷纷加快ASIC芯片的自研和测试；国内受AI芯片禁令影响，英伟达先进AI芯片对华出口受限，互联网大厂可能转向AI ASIC芯片，服务器厂商有望充分受益；同时，国产算力芯片25年有望快速放量，建议关注海光信息等。同时，香港《稳定币条例》落地，稳定币有望提升跨境支付效率，建议关注新大陆等公司。

风险提示：互联网大厂AI ASIC研发进展不及预期；云厂商资本开支投入不及预期；稳定币发行进展不及预期。

重点公司盈利预测及投资评级

公司代码	公司名称	投资评级	昨收盘（元）	总市值（百万元）	EPS 2025E	EPS 2026E	PE 2025E	PE 2026E
688041	海光信息	优于大市	136.10	316,342	1.69	2.36	80.53	57.67
000997	新大陆	优于大市	32.71	33,759	1.18	1.42	27.72	23.04

资料来源：Wind、国信证券经济研究所预测

行业研究·行业月报

计算机

优于大市·维持

证券分析师：熊莉
021-61761067
xiongli1@guosen.com.cn
S0980519030002

证券分析师：库宏垚
021-60875168
kuhongyao@guosen.com.cn
S0980520010001

证券分析师：艾宪
0755-22941051
aixian@guosen.com.cn
S0980524090001

联系人：云梦泽
021-60933155
yunmengze@guosen.com.cn

联系人：蔺亚婴
linyaying@guosen.com.cn

联系人：侯睿
hourui3@guosen.com.cn

市场走势

资料来源：Wind、国信证券经济研究所整理

相关研究报告

《计算机行业2025年6月暨中期投资策略-AI产业快速迭代,持续看好Agent和算力租赁》 ——2025-06-13
《稳定币香港政策落地，关注板块投资机会》 ——2025-06-04
《人工智能专题报告：国内大厂扩张资本开支，算力租赁订单持续落地》 ——2025-05-21
《计算机行业2025年5月投资策略暨财报总结-大厂布局Agent产品，AI应用快速落地》 ——2025-05-08
《人工智能行业专题：2025Q1海外大厂CapEx和ROIC总结梳理-20250505》 ——2025-05-05

内容目录

AI ASIC 芯片：市场规模快速增长，三大发展趋势逐渐明朗 5
AI Asic 芯片：价格、功率优势明显，市场规模快速增长 5
复盘谷歌 TPU：更专用、强算力、大集群、高能效 8
稳定币：香港政策落地，关注板块投资机会 17
香港《稳定币条例》落地，以合规框架激活数字金融创新 17
稳定币蓬勃发展，跨境支付+RWA 有望收益 18
投资建议:看好 AI ASIC 及稳定币 20
风险提示 20

图表目录

图1： 不同类型 AI 芯片对比 5
图2： GPU 和 AI ASIC 平均单价及预测 6
图3： AI 芯片算力和功率矩阵图 6
图4： 24Q4 各厂商 AI 芯片收入占比 7
图5： 推理、训练侧芯片类型 TAM（2024vs2030） 7
图6： GPU、AI ASIC 芯片市场规模情况（单位：十亿美金） 7
图7： GPU、AI ASIC 芯片出货量情况（单位：万颗） 7
图8： 谷歌 TPU 发展历史 8
图9： TPU v1 Floor Plan 9
图10： TPU v1 Block Diagram 9
图11： TPU v1 向 TPU v2 架构演进 9
图12： TPU v2 芯片包含两个相连的 Tensor Core 10
图13： TPU v2 Floorplan 10
图14： TPU v3 延续 v2 架构，性能提升 10
图15： TPU v4 MXU 数量翻倍，峰值算力大幅提升 11
图16： 谷歌超级计算机互联结构（Cube） 11
图17： 可重配置光互连技术提升计算机的稳定性 11
图18： 谷歌 TPU v4 性能表现优于英伟达 A100 12
图19： 谷歌 TPU v4 性能表现略逊于 H100，但功耗管理能力出色 12
图20： 谷歌 TPU v5e 架构 12
图21： 谷歌 TPU v4、TPU v5e、TPU v5p 参数对比 12
图22： 谷歌 TPU v5e 单美元推理性能性价比提升 13
图23： 谷歌 TPU v5e 延迟相较于 TPU v4 进一步下降 13
图24： TPU v5e Pod 可承载 2 万亿模型运行 13
图25： TPU v6e（Trillium）同 TPU v5e 参数对比 14
图26： 谷歌 Trillium 实现 99%的拓展效率（12 个 Pod） 14
图27： 谷歌 Trillum 同 TPU v5p 拓展效率对比 14
图28： 谷歌 Trillium 训练效率对比（同 TPU v5e） 14
图29： 谷歌 Trillium 对 MoE 架构模型训练能力大幅提升 14
图30： 谷歌 Trillium 推理效率对比（同 TPU v5e） 15
图31： 谷歌 Trillium 推理性价比对比（同 TPU v5e） 15
图32： TPUv4、TPUv5p、Ironwood 芯片参数对比 15
图33： Ironwood（TPU v7）峰值性能大幅提升 16
图34： Ironwoord（TPU v7）峰值能效大幅提升 16
图35： 各类稳定币占比 17
图36： USDT 与 USDC 对照表 17
图37： 《稳定币条例》收益类型、政策要点及代表企业 18

图38： 稳定币利好“沙盒”参与者、跨境支付服务提供商、RWA 项目方三种类型主体 18
图39： 稳定币支付缩短跨境交易结算时间 19
图40： SWIFT 系统按货币统计平均汇款处理时间 19
图41： USDT 发行和流通过程 19
图42： 稳定币全年每天 24 小时不间断交易 19
图43： RWA 项目实践 20

AI ASIC 芯片：市场规模快速增长，三大发展趋势逐渐明朗

AI Asic 芯片：价格、功率优势明显，市场规模快速增长

AI 芯片分类：AI 芯片指专门用于运行人工智能算法且做了优化设计的芯片，为满足不同场景下的人工智能应用需求，AI 芯片逐渐表现出专用性、多样性的特点。根据设计需求，AI 芯片主要分为中央处理器（CPU）、图形处理器（GPU）、现场可编程逻辑门阵列（FPGA）、专用集成电路（ASIC）等，相比于其他 AI 芯片，ASIC 具有性能高、体积小、功率低等特点。

CPU->GPU->ASIC，ASIC 成为 AI 芯片重要分支。1）CPU 阶段：尚未出现突破性的 AI 算法，且能获取的数据较为有限，传统 CPU 可满足算力要求；2）GPU 阶段：2006 年英伟达发布 CUDA 架构，第一次让 GPU 具备了可编程性，GPU 开始大规模应用于 AI 领域；3）ASIC 阶段：2016 年，Google 发布 TPU 芯片（ASIC 类），ASIC 克服了 GPU 价格昂贵、功耗高的缺点，ASIC 芯片开始逐步应用于 AI 领域，成为 AI 芯片的重要分支。

图1：不同类型 AI 芯片对比

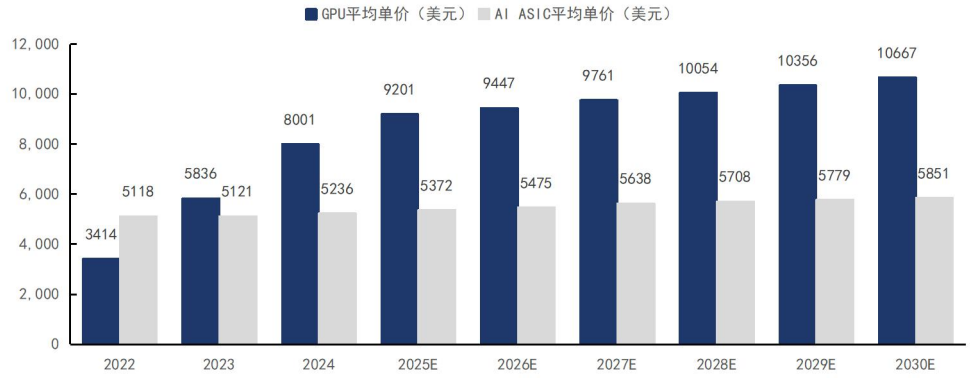
类别	CPU	GPU	FPGA	ASIC
特点	拥有大量的缓存和复杂的逻辑控制单元	一种由大量运算单元组成的大规模并行计算架构芯片	可对其集成的基本门电路和存储器进行重新定义	全定制化芯片，其无法通过修改电路进行功能拓展
功耗	高	高	中	低
优势	✓ 灵活性 ✓ 通用性强 ✓ 复杂指令和任务 ✓ 系统管理	✓ 大量并行核 ✓ AI处理出色表现	✓ 可配置的逻辑门 ✓ 灵活性 ✓ 可重新编程性	✓ 可用库设计的定制化逻辑 ✓ 更快的处理速度 ✓ 体积小
劣势	✓ 核数少 ✓ 时延严重 ✓ 效率低	✓ 功耗高 ✓ 体积大	✓ 编程复杂	✓ 固定的功能 ✓ 前期定制化成本高
代表厂商	Intel、AMD	NVIDIA、AMD	Xilinx、Altera	Google、寒武纪
	 Intel Sapphire Rapids	 NVIDIA H100	 Xilinx Versal AI Core	 Google TPU

资料来源：Ashutosh Mishra 等著-《Artificial Intelligence and Hardware Accelerators》-2023 年 Springer 出版-P35，国信证券经济研究所整理

优势一：相比于 GPU 算力卡，AI ASIC 芯片价格优势明显

由于 GPU 芯片强大的通用性、灵活性，其设计及流片成本较高，进而导致 GPU 的平均单价较高。从历史趋势来看，根据 IDC 统计数据，2022-2024 年受 AI 大模型驱动，GPU 性能需求快速提升，进而导致 GPU 产品的平均单价快速提升（对应 22-24 年 CAGR 为 53.1%），从短期来看，2024 年 GPU 平均单价为 8001 美元，AI ASIC 平均单价为 5236 美元，AI ASIC 具备价格优势。从长期来看，根据 IDC 预测数据，GPU 平均单价自 2025 年后稳中有升，AI ASIC 平均单价基本维稳，预计 2030 年 GPU 和 AI ASIC 平均单价分别为 10667、5851 美元，AI ASIC 价格优势仍然明显。

图2：GPU 和 AI ASIC 平均单价及预测

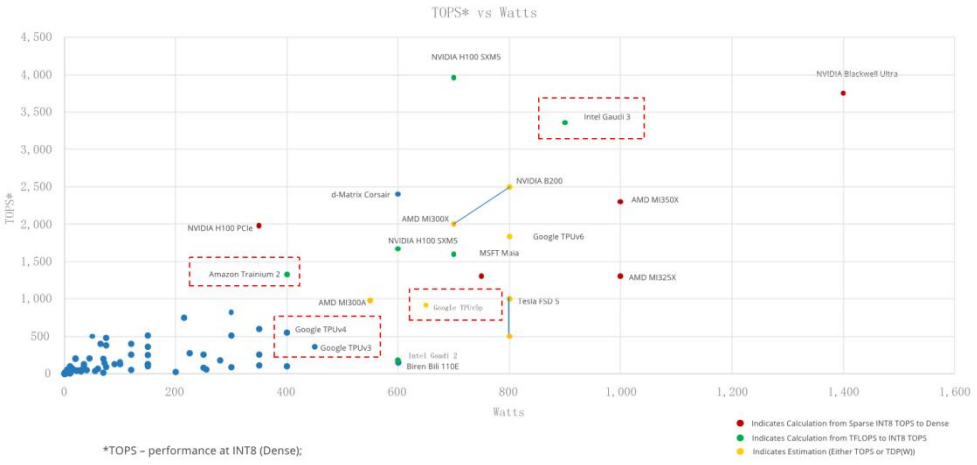


资料来源：IDC，国信证券经济研究所整理

优势二：相比于 GPU 算力卡，AI ASIC 芯片功率更低，能耗优势明显

由于 AI AISC 芯片偏定制化设计，专为特定任务（例如 AI 大模型的训练或推理）优化，因此其在执行特定任务时功率较低。根据 IDC 统计数据，在同等算力水平下，AI ASIC 的功率更低，能耗优势明显（例如 Amazon 的 ASIC 芯片 Trainium 2 同 AMD 的 MI300A 比较，谷歌 ASIC 芯片 TPU v6 同 AMD 的 MI325 比较，Intel 的 Gaudi3 同英伟达的 Blackwll Ultra 比较）。

图3：AI 芯片算力和功率矩阵图



资料来源：IDC，国信证券经济研究所整理

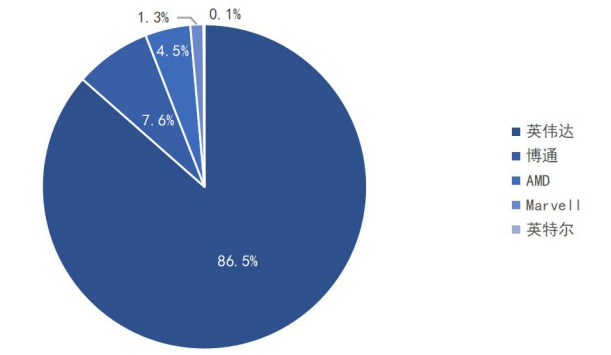
AI ASIC 市场规模：2024 年 148 亿美金，预计 2030 年增长至 838 亿美金

ASIC 芯片在训练、训练&推理双用 AI 芯片领域，增速快于 GPU。根据 IDC 披露数据，24Q4 英伟达、博通、AMD、Marvell、英特尔 AI 芯片收入占比分别为 86.5%、7.6%、4.5%、1.3%、0.1%，其中英伟达和 AMD 为 GPU 算力卡，博通、Marvell 分

别为谷歌、亚马逊定制 ASIC 芯片，英特尔为自研 ASIC 芯片，对博通、Marvell、英特尔市占率进行加总，则 24Q4 AI ASIC 芯片占比为 9.0%。分领域来看：

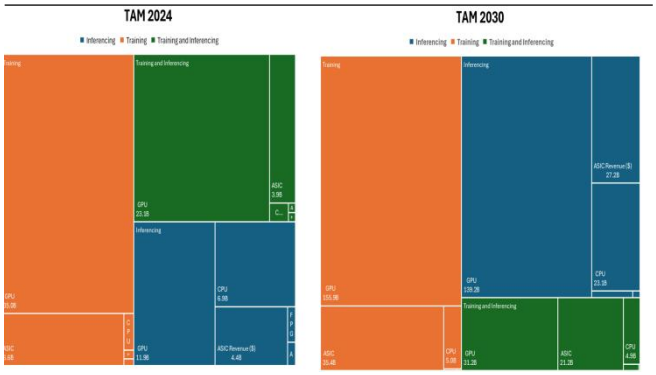
- **训练用 AI 芯片：**2024 年训练用 AI 芯片 TAM 中，GPU、ASIC 分别为 350、66 亿美金，预计 2030 分别提升至 1559、354 亿美金，对应 24-30 年 CAGR 为 28.3%、32.3%。
- **推理用 AI 芯片：**2024 年推理用 AI 芯片 TAM 中，GPU、ASIC 分别为 119、44 亿美金，预计 2030 分别提升至 1392、272 亿美金，对应 24-30 年 CAGR 为 50.7%、35.5%。
- **训练&推理双用 AI 芯片：**2024 年训练&推理双用 AI 芯片 TAM 中，GPU、ASIC 分别为 231、39 亿美金，预计 2030 分别提升至 312、212 亿美金，对应 24-30 年 CAGR 为 5.1%、32.6%。

图4：24Q4 各厂商 AI 芯片收入占比



资料来源：IDC，国信证券经济研究所整理

图5：推理、训练侧芯片类型 TAM（2024vs2030）



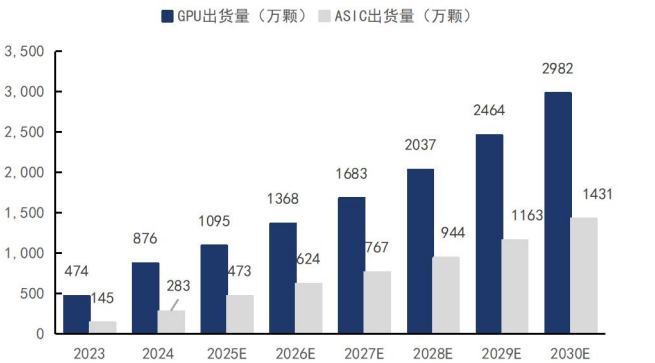
资料来源：IDC，国信证券经济研究所整理

AI ASIC 芯片市场规模、出货量快速增长。从市场规模来看，根据 IDC 披露数据，2024 年 GPU、AI ASIC 芯片市场规模分别为 701、148 亿美金，预计 2030 年分别增长至 3263、838 亿美金，对应 24-30 年 CAGR 分别为 29.2%、33.5%。从出货量来看，2024 年 GPU、AI ASIC 芯片出货量分别为 876、283 万颗，预计 2030 年增长至 2982、1431 万颗，对应 24-30 年 CAGR 分别为 22.6%、31.0%，AI ASIC 芯片占比稳步提升。

图6：GPU、AI ASIC 芯片市场规模情况（单位：十亿美金）

资料来源：IDC，国信证券经济研究所整理

图7：GPU、AI ASIC 芯片出货量情况（单位：万颗）

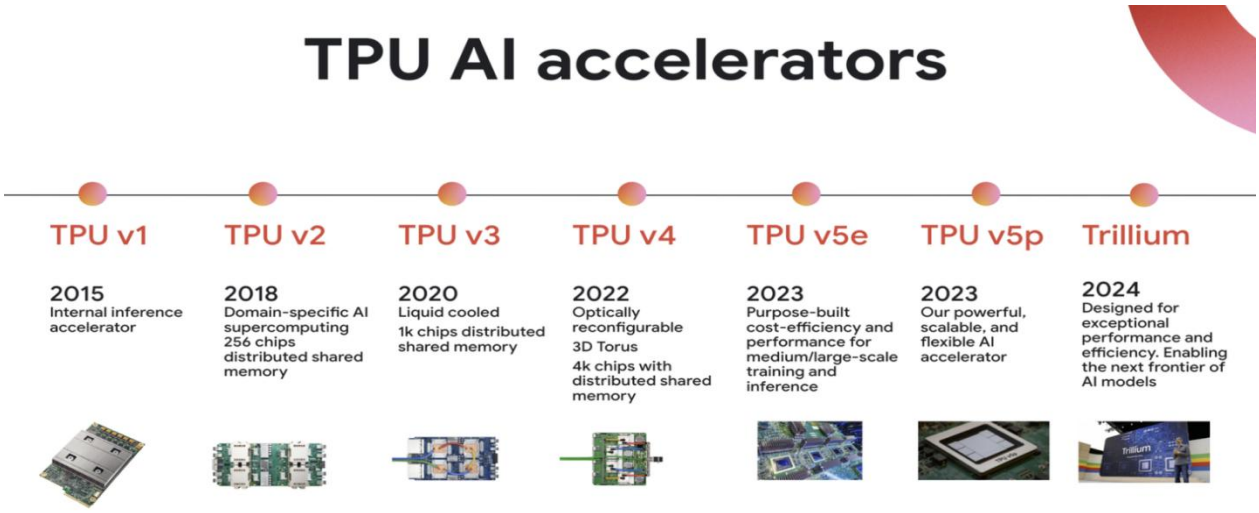


资料来源：IDC，国信证券经济研究所整理

复盘谷歌 TPU：更专用、强算力、大集群、高能效

谷歌 TPU 发展历程：2015 年发布 TPU v1，与使用通用 CPU 和 GPU 的神经网络计算相比，TPU v1 带来了 15-30 倍的性能提升和 30-80 倍的能效提升，其以较低成本支持谷歌的很多服务，仅可用于推理；18 年发布 TPU v2，用于加速大量的机器学习和人工智能工作负载，包括训练和推理；20 年发布 TPU v3，算力和功率大幅增长，其采用了当时最新的液冷技术；22 年发布 TPU v4（包括 TPU v4i），应用 7nm 工艺，晶体管数大幅提升，算力提升，功耗下降；23 年发布 TPU v5e 和 TPU v5p，其中 TPU v5e 专为提升大中型模型的训练、推理性能以及成本效益所设计，使企业能够以更低的成本，训练和部署更大、更复杂的 AI 模型；24 年发布第六代 TPU——Trillium，训练、推理性能和能效比大幅提升，首次加入了专为 Transformer 类大语言模型优化的大规模 MLP(多层感知器)核心，与标准 TPU 核心协同工作，将进一步提升大模型的训练速度与效率，同时发布了基于 Trillium 芯片的全新机架系统 TPU v6 Pod，以满足大规模集群部署的需要；25 年谷歌发布第七代 TPU——Ironwood，首款在其张量核和矩阵数学单元中支持 FP8 计算，同时 HBM 容量大幅提升，可处理更大型的模型和数据集运算。

图8：谷歌 TPU 发展历史

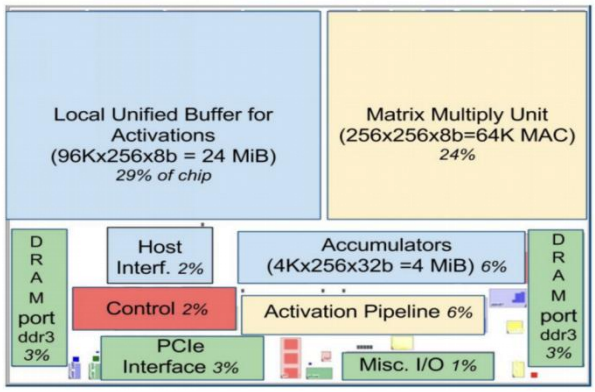


资料来源：谷歌云，国信证券经济研究所整理

谷歌 TPU v1：谷歌第一代 TPU 产品，仅可用于推理。TPU v1 主要包括统一缓冲器（Unified Buffer）、矩阵乘法单元（MMU）、累加器（Accumulators）、激活流水线电路（Activation Pipeline）、DDAM 等，其中统一缓冲器和矩阵乘法单元面积占比最高，合计达 53%。

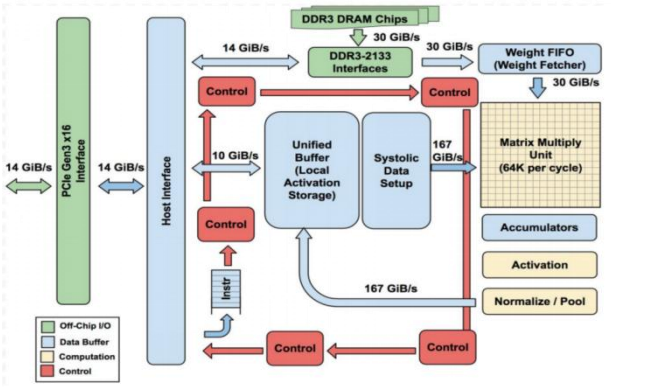
TPU v1 工作流程：1）芯片启动，缓冲区和 DDR3 为空；2）用户加载 TPU 编译的模型，将权重放入 DDR3 内存；3）主机用输入值填充激活缓冲区；4）发送控制信号将一层权重加载到矩阵乘法单元；5）主机触发执行，激活并通过矩阵乘法单元传播到累加器；6）通过激活流水线电路，新层替换缓冲区的旧层；7）重复步骤 4-7，直到最后一层；8）最后一层的激活被发送给主机。

图9：TPU v1 Floor Plan



资料来源：Norman P. J 等-《In-Datacenter Performance Analysis of a Tensor Processing Unit》-ISCA（2017）-P3，国信证券经济研究所整理

图10：TPU v1 Block Diagram

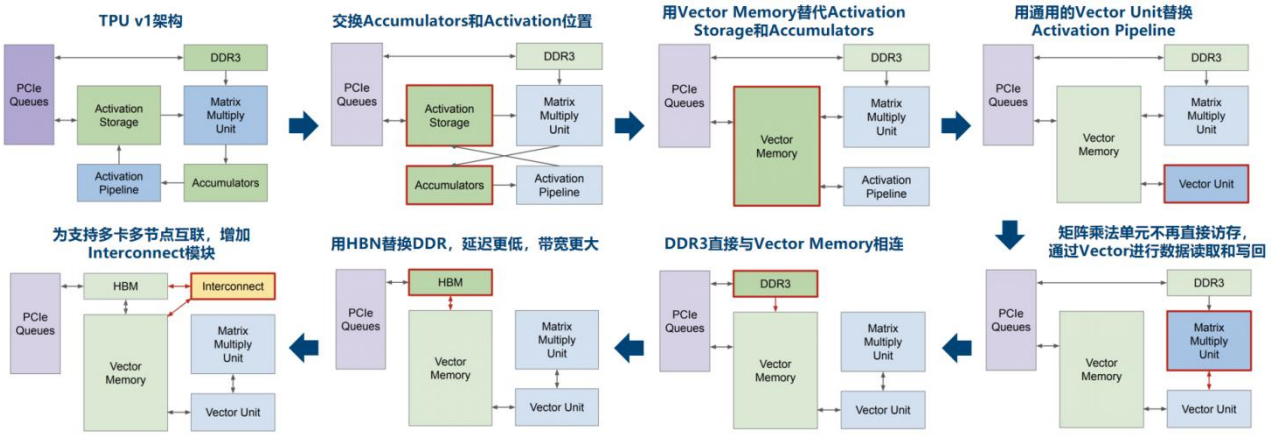


资料来源：Norman P. J 等-《In-Datacenter Performance Analysis of a Tensor Processing Unit》-ISCA（2017）-P3，国信证券经济研究所整理

谷歌 TPU v2：架构大规模更新，增加训练功能。谷歌 TPU v2 是基于 TPU v1 架构的大规模更新，使其在推理的基础上增加训练功能，主要体现为以下三点：

- **更大的灵活性：**训练面对不同算法，需要更大的灵活性，TPU v2 将 Activation Storage 和 Accumulators 两个相互独立的缓冲区合并成一个 Vector Memory，进而提高可编程性；添加了可编程更高 Vector Unit，用于替代固定的 Activation Pipeline。
- **更大的内存：**训练既需要读取权重，也要写入权重，所以将 DDR3 直接与 Vector Memory 相连，并用 HBM 替代 DDR3，延迟更低，带宽更大。
- **提供拓展能力（集群方案）：**为了加速训练，通常会采用集群方案，添加 Interconnect 可以使其与其他 TPU 进行高效互换。

图11：TPU v1 向 TPU v2 架构演进

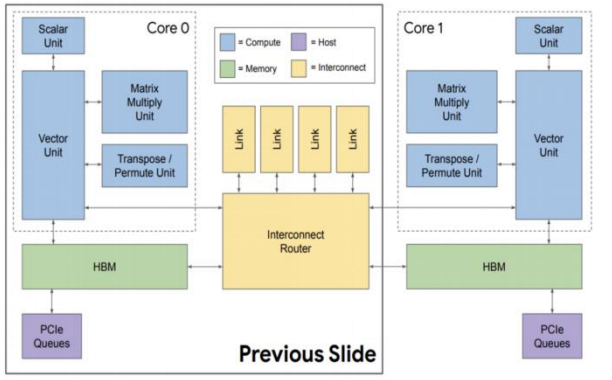


资料来源：Norrie T 等-《The Design Process for Google's Training Chips: TPUv2 and TPUv3》-IEEE（2020）-P3，国信证券经济研究所整理

同时，TPU v2 的内核数量和 MXU 利用率进一步提升。1）内核数量：TPU v1 仅有 1 个 Tensor Core，导致管道更为冗长；TPU v2 的内核数增加为 2 个，对编译器也更为友好。**2）MXU 利用率：**TPU v1 的 MXU 包含 256*256 个乘积累加运算器，由于部分卷积计算规模小于 256*256，导致单个大核的利用率相对较低；而 TPU v2

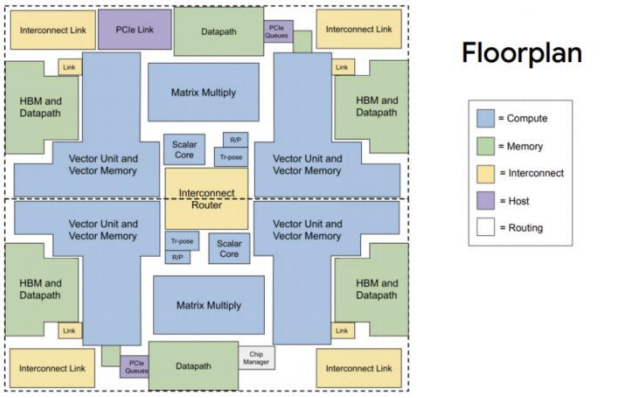
的单核 MXU 包含 128*128 个乘积累加运算器，在一定程度上，提升了 MXU 利用率。

图12：TPU v2 芯片包含两个相连的 Tensor Core



资料来源：Norrie T 等-《The Design Process for Google’s Training Chips: TPUv2 and TPUv3》-IEEE（2020）-P3，国信证券经济研究所整理

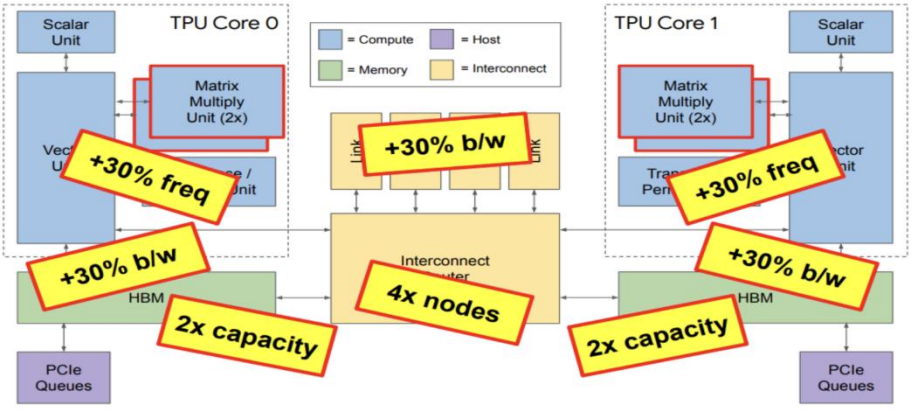
图13：TPU v2 Floorplan



资料来源：Norrie T 等-《The Design Process for Google’s Training Chips: TPUv2 and TPUv3》-IEEE（2020）-P7，国信证券经济研究所整理

谷歌 TPU v3：延续 v2 架构，性能提升，首次采用液冷。TPU V3 在 v2 架构的基础上，矩阵乘法单元（MXU）数量提升翻倍，时钟频率加快 30%，内存带宽加大 30%，HBM 容量翻倍，芯片间带宽扩大了 30%，可连接的节点数为先前 4 倍，性能大幅提升；同时，首次采用液冷技术，峰值算力为 TPU v2 的 2.67 倍，而 TDP 仅为 TPU v2 的 1.61 倍，TDP 大幅优化。

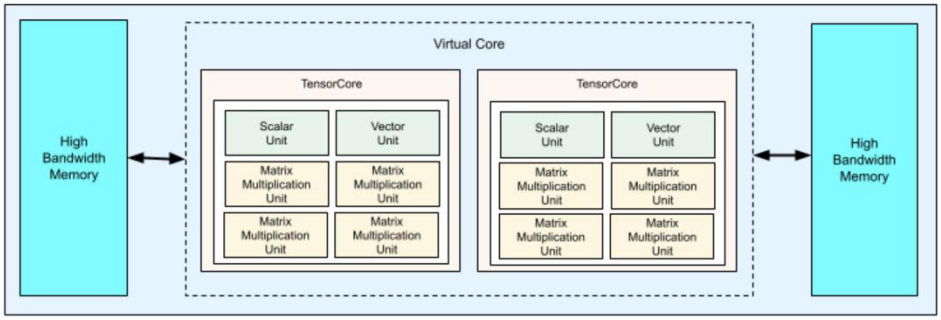
图14：TPU v3 延续 v2 架构，性能提升



资料来源：Norrie T 等-《Google’s Training Chips Revealed: TPUv2 and TPUv3》-IEEE（2020）-P49，国信证券经济研究所整理

谷歌 TPU v4：采用 7nm 生产工艺，性能大幅提升。从硬件提升来看，根据 Google Cloud 数据，TPU v4 芯片包含 2 个 TensorCore，每个 TensorCore 包含 4 个 MXU，是 TPU v3 的 2 倍；同时，HBM 带宽提升至 1200 GBps，相比上一代，提升 33.33%。从峰值算力来看，TPU v4 的峰值算力达 275 TFLOPS，为 TPU v3 峰值算力的 2.24 倍。

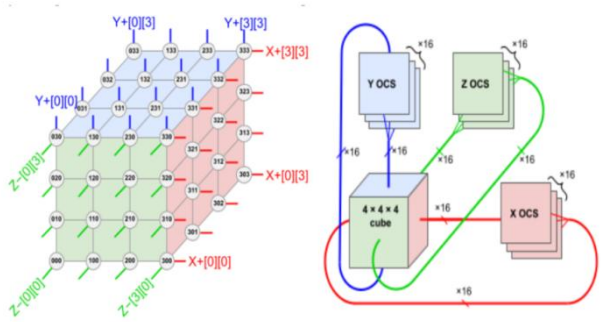
图15：TPU v4 MXU 数量翻倍，峰值算力大幅提升



资料来源：谷歌云，国信证券经济研究所整理

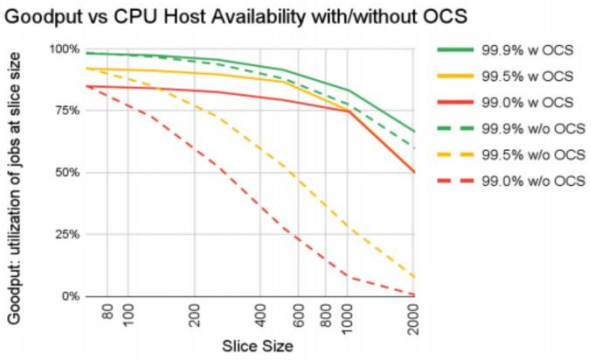
谷歌基于 TPU v4 提出 Cube 互联架构（3D 拓扑架构），发布可重配置光互连技术。谷歌提出将 4*4*4（64）个 TPU v4 芯片连接成 1 个立方体结构（Cube），再将 4*4*4 个立方体结构（Cube）连接成共有 4096 个 TPU v4 芯片的超级计算机，其中物理距离较近 TPU v4 芯片（即同一个 Cube 中的 4*4*4 个芯片）采用常规电互联方式，距离较远的 TPU（例如 Cube 之间的互联）间用光互连。采用光互连技术可以有效避免“芯片等数据”的情形出现，进而提升计算效率。同时，TPU v4 通过加入光路开关（OCS）的方式，可以根据具体模型数据流来调整 TPU 之间的互联拓扑，实现最优性能，根据《TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings》数据，可重配置光互连技术可以将性能提升至先前的 1.2-2.3 倍。

图16：谷歌超级计算机互联结构（Cube）



资料来源：Norman P. J 等-《TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings》-ISCA（2023）-P2，国信证券经济研究所整理

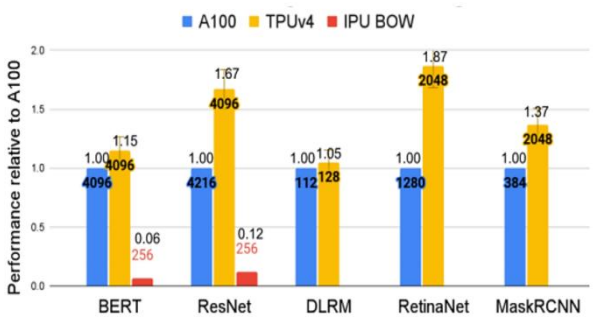
图17：可重配置光互连技术提升计算机的稳定性



资料来源：Norman P. J 等-《TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings》-ISCA（2023）-P3，国信证券经济研究所整理

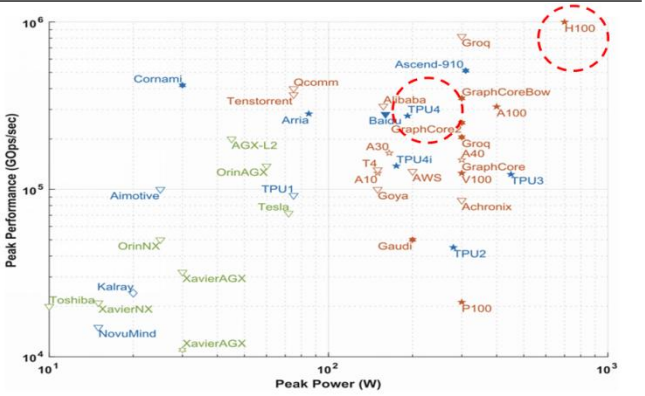
谷歌 TPU v4 性能介于英伟达 A100 和 H100 之间。根据《TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings》数据，TPU v4 的性能表现在 BERT、ResNet、DLRM、RetinaNet、MaskRCNN 下分别为 A100 的 1.15x、1.67x、1.05x、1.87x 和 1.37x，性能表现优于英伟达 A100。同时，根据《AI and ML Accelerator Survey and Trends》数据，英伟达 H100 的峰值性能表现高于 TUP v4，而 TUP v4 作为 ASIC 芯片，在功耗管理方面表现出色，峰值功率低于 H100。

图18：谷歌 TPU v4 性能表现优于英伟达 A100



资料来源：Norman P. J 等-《TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings》-ISCA（2023）-P9，国信证券经济研究所整理

图19：谷歌 TPU v4 性能表现略逊于 H100，但功耗管理能力出色

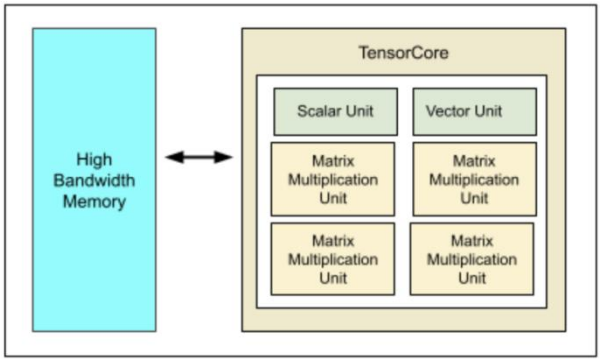


资料来源：Reuther A 等-《AI and ML Accelerator Survey and Trends》-IEEE（2022）-P3，国信证券经济研究所整理

谷歌 TPU v5：分为 TPU v5e 和 TPU v5p 两个版本。

- **TPU v5e：训推一体，强调成本效益“cost-efficient”和可拓展性。**从芯片架构来看，TPU v5e 采用单 TensorCore 架构，每个 TensorCore 包含 4 个矩阵乘法单元（MXU）、1 个向量单元和 1 个标量单元，基本延续了上一代的架构。**1）成本效益：**TPU v5e 将 HBM 显存和带宽降低（16GB 显存基本可以承载百亿参数模型推理任务），且采用单 TensorCore 架构，成本大幅下降，同时提升 INT 8 精度算力，结合软硬件优化，TPU v5e 实现了每美元推理性价比提升高达 2.5 倍，推理延迟降低 1.7 倍；**2）可拓展性：**根据谷歌云披露数据，单颗 TPU v5e 芯片可以运行高达 130 亿模型，可以拓展至 256 颗芯片，运行 2 万亿参数大模型。
- **TPU v5p：专注于超大基础模型训练，算力、HBM、Pod 规模大幅提升，AI 模型训练速度、性价比表现出色。1）模型训练速度：**同 TPU v4 相比，单一 Pod 芯片数量从 4096 颗提升至 8960 颗，单卡算力、HBM 显存是 TPU v4 的 2 倍、3 倍，根据谷歌云披露数据，TPU v5p 训练速度是 TPU v4 的 2.8 倍（以 1750 亿的 GPT-3 为例）；**2）性价比：**根据谷歌云披露数据，以 GPT-3 为例，TPU v5p 每美元芯片表现性价比是 TPU v4 的 2.1 倍。

图20：谷歌 TPU v5e 架构



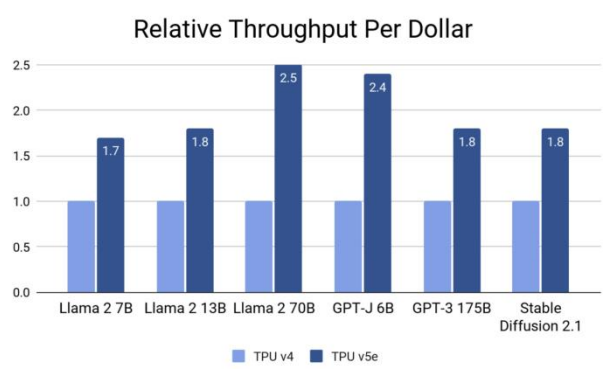
资料来源：谷歌云，国信证券经济研究所整理

图21：谷歌 TPU v4、TPU v5e、TPU v5p 参数对比

	v4	v5e	v5p
Chips per pod	4096	256	8,960
Chip Bf16 TFLOPs	275	197	459
Chip Int8 TOPs	N/A	394	918
HBM (GB)	32	16	95
HBM BW (GB/s)	1228	820	2,765
ICI BW per chip (Gb/s)	2,400	1,600	4,800

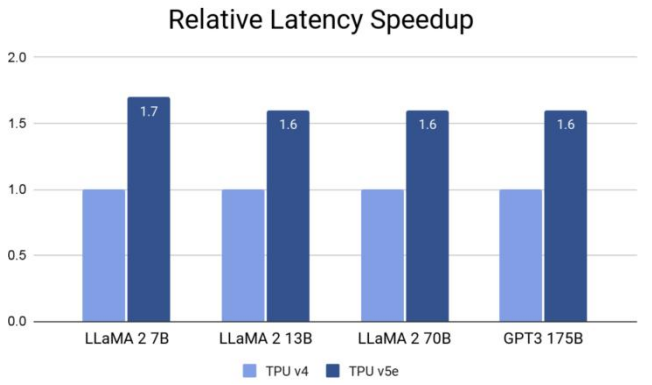
资料来源：谷歌云，国信证券经济研究所整理

图22：谷歌 TPU v5e 单美元推理性能性价比提升



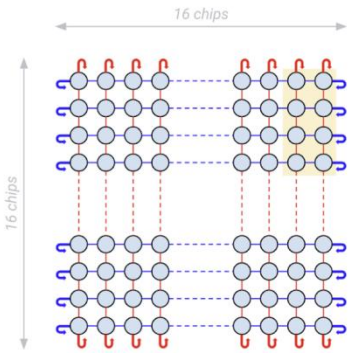
资料来源：谷歌云，国信证券经济研究所整理

图23：谷歌 TPU v5e 延迟相较于 TPU v4 进一步下降



资料来源：谷歌云，国信证券经济研究所整理

图24：TPU v5e Pod 可承载 2 万亿模型运行



TPU v5e Chips	Max Model Size* (# parameters)	Reference Models & Sizes
1	13 Billion	LLaMA 2 13B
4	32.5 Billion	LLaMA 32.5B
8	65 Billion	LLaMA 65B
16	175 Billion	GPT-3 175B
32	280 Billion	Gopher 280B
64	540 Billion	PaLM 540B
128	1 Trillion	GLaM 1T
256 (1 Pod)	2 Trillion	Switch Transformer 1.6T

资料来源：谷歌云，国信证券经济研究所整理

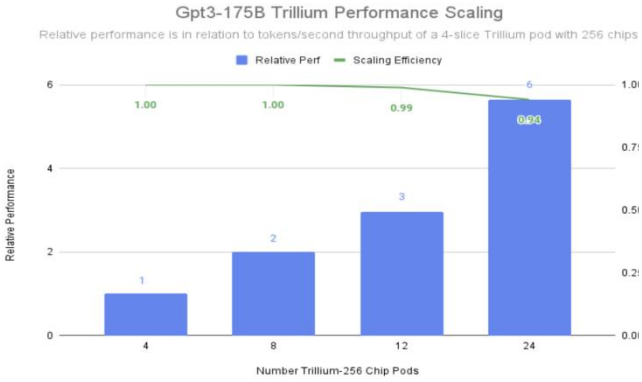
谷歌 TPU v6：发布第六代 TPU Trillium（TPU v6e），拥有接近线性的拓展能力，训练、推理性能大幅提升。Trillium（TPU v6e）产品定位同 TPU v5e 相同，训推一体算力卡，硬件架构延续了 TPU v5e，在各精度算力、HBM 等维度做了大幅提升。**1）接近线性的拓展能力：**Trillium 通过高速芯片互联、Jupiter 网络连接，实现在众多 Trillium 主机上高效地分配工作负载，根据谷歌云披露数据，3072 个 Trillium（对应 12 个 Pod）对 1750 亿的 GPT-3 模型进行预训练，拓展效率达 99%。**2）训练、推理性能大幅提升：**根据谷歌云披露数据，相较于 TPU v5e，Trillium 对 GPT-3（1750 亿参数）、Llama-2（700 亿参数）等密集 LLM 的训练速度分别提升 3.24、4.0 倍，同时优化对 MoE 架构模型的训练能力；此外，与 TPU v5e 相比，Trillium 在 Stable Diffusion XL 上的离线推理相对吞吐量（每秒图像数）提高了 3.1 倍，服务器推理相对吞吐量提高了 2.9 倍。**3）性比价持续提升：**根据谷歌云披露数据，在 Trillium 上生成一千张图像的成本，离线推理比 TPU v5e 低 27%，服务器推理 SDXL 比 TPU v5e 低 22%。

图25：TPU v6e（Trillium）同 TPU v5e 参数对比

Specification	v5e	v6e
Performance/total cost of ownership (TCO) (expected)	0.65x	1
Peak compute per chip (bf16)	197 TFLOPs	918 TFLOPs
Peak compute per chip (Int8)	393 TOPs	1836 TOPs
HBM capacity per chip	16 GB	32 GB
HBM bandwidth per chip	819 GBps	1640 GBps
Inter-chip interconnect (ICI) bandwidth	1600 Gbps	3584 Gbps
ICI ports per chip	4	4
DRAM per host	512 GiB	1536 GiB
Chips per host	8	8
TPU Pod size	256 chips	256 chips
Interconnect topology	2D torus	2D torus
BF16 peak compute per Pod	50.63 PFLOPs	234.9 PFLOPs
All-reduce bandwidth per Pod	51.2 TB/s	102.4 TB/s
Bisection bandwidth per Pod	1.6 TB/s	3.2 TB/s
Per-host NIC configuration	2 x 100 Gbps NIC	4 x 200 Gbps NIC
Data center network bandwidth per Pod	6.4 Tbps	25.6 Tbps
Special features	-	SparseCore

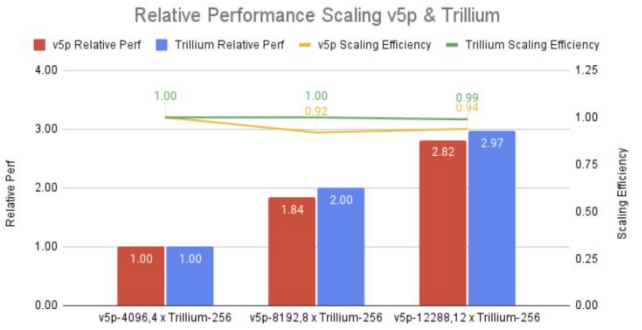
资料来源：谷歌云，国信证券经济研究所整理

图26：谷歌 Trillium 实现 99%的拓展效率（12 个 Pod）



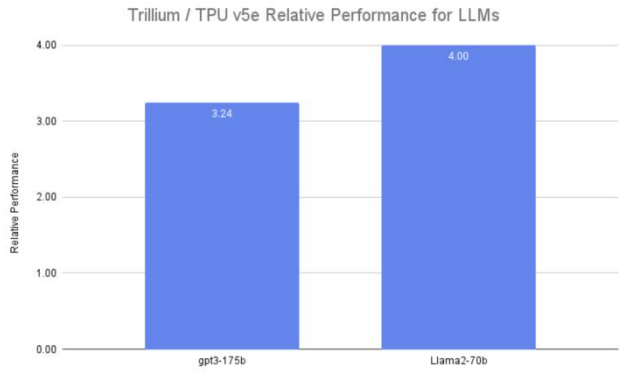
资料来源：谷歌云，国信证券经济研究所整理

图27：谷歌 Trillum 同 TPU v5p 拓展效率对比



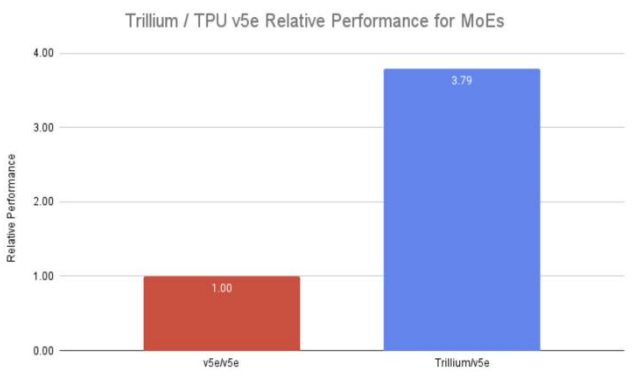
资料来源：谷歌云，国信证券经济研究所整理

图28：谷歌 Trillium 训练效率对比（同 TPU v5e）



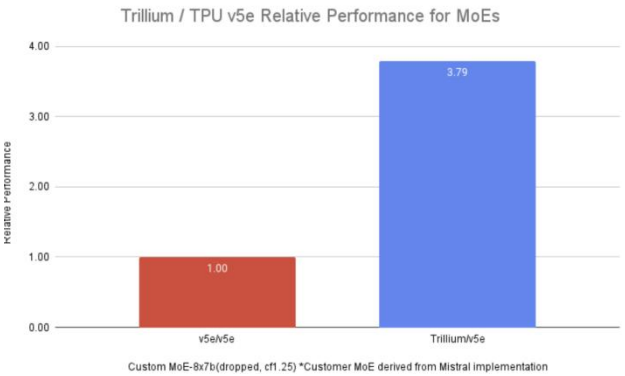
资料来源：谷歌云，国信证券经济研究所整理

图29：谷歌 Trillium 对 MoE 架构模型训练能力大幅提升



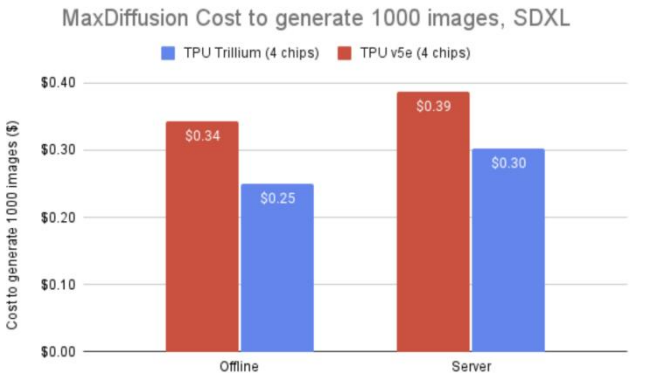
资料来源：谷歌云，国信证券经济研究所整理

图30：谷歌 Trillium 推理效率对比（同 TPU v5e）



资料来源：谷歌云，国信证券经济研究所整理

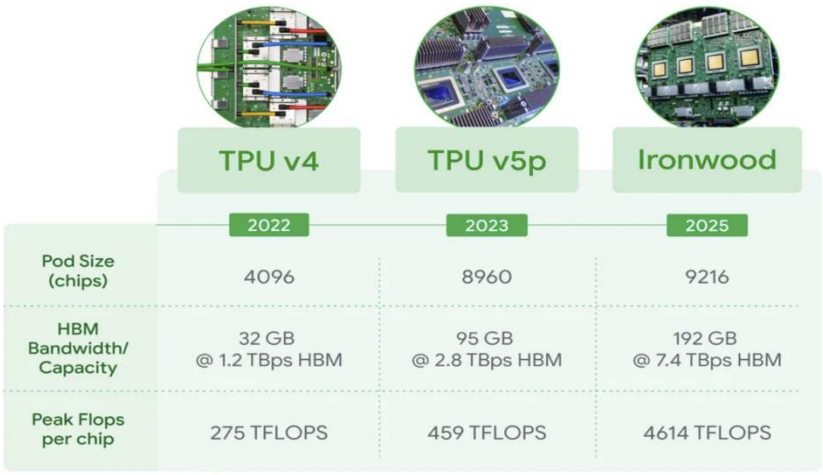
图31：谷歌 Trillium 推理性价比对比（同 TPU v5e）



资料来源：谷歌云，国信证券经济研究所整理

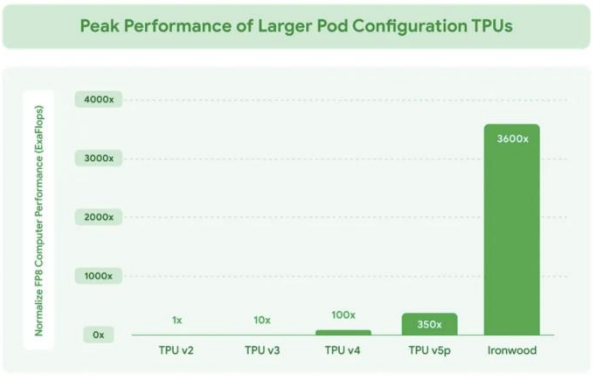
谷歌 TPU v7：性能大幅提升，单 Pod 规模进一步扩大，能效表现优秀。1）性能大幅提升，单 Pod 规模进一步扩大：Ironwood（TPU v7）产品定位同 TPU v5p 相同，为大规模的思考型、推理型 AI 模型提供动力，单芯片峰值 Flops 达 4614 TFLOPS，约为 TPU v5p 的 10x，HBM 容量、芯片间互联（ICI）带宽分别提升至 192GB、1.2 TBps，单 Pod 尺寸进一步拓展至 9216 颗芯片，峰值性能大幅提升；**2）能耗表现优秀：**根据谷歌云披露数据，以单芯片封装每瓦热设计功耗所提供的峰值 FP8 Flops 衡量，Ironwood 峰值能效是上一代 Trillium 的 2 倍，是 TPU v2 的 29.3 倍。

图32：TPUv4、TPUv5p、Ironwood 芯片参数对比



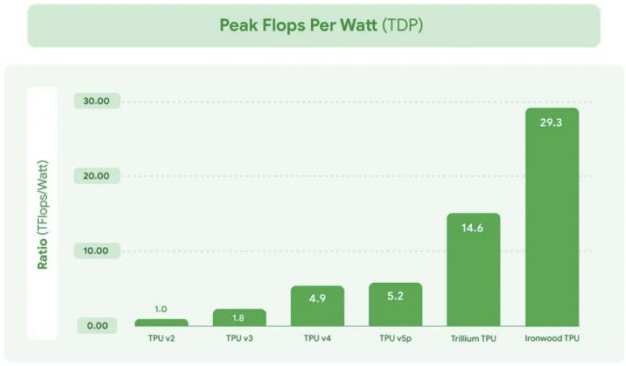
资料来源：谷歌云，国信证券经济研究所整理

图33：Ironwood（TPU v7）峰值性能大幅提升



资料来源：谷歌云，国信证券经济研究所整理

图34：Ironwoord（TPU v7）峰值能效大幅提升



资料来源：谷歌云，国信证券经济研究所整理

通过复盘谷歌 AI ASIC 芯片发展历史，AI ASIC 芯片发展呈现出三个趋势：

- **趋势一：专用性持续增强，颗粒度更细**。最初的 TPU v1 仅支持推理任务，TPU v2 可以支持训练和推理任务，TPU 发展的前期是拓展芯片能力；2022 年谷歌发布 TPU v4 时，同时发布 TPU v4i，其中 TPU v4i 属于 lite 版本，适用于推理任务。TPU v5 分为了 TPU v5e 和 TPU v5p 两个版本，其中 TPU v5e 是训推一体，强调成本效益，而 TPU v5p 性能强劲，专注于超大基础模型训练；后续 Trillium、Ironwood 分别走 TPU v5e、TPU v5p 路线，芯片应用场景更细分，专用性显著。
- **趋势二：更强的算力、HBM 和集群能力**。AI ASIC 芯片的单卡算力持续提升，选用更领先的 HBM（应对多模态任务），单 POD 芯片数量持续提升，集群拓展效率逐步接近线性。
- **趋势三：能效比持续提升**。基于智算中心电力消耗巨大，AI ASIC 芯片的能效持续提升，例如以单芯片封装每瓦热设计功耗所提供的峰值 FP8 Flops 衡量，Ironwood 峰值能效是上一代 Trillium 的 2 倍，是 TPU v2 的 29.3 倍；同时，TPU v3 开始配套液冷，液冷等新一代冷却方式逐步应用。

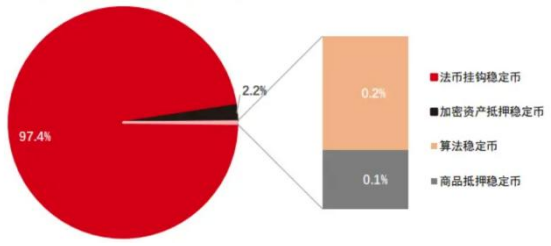
稳定币：香港政策落地，关注板块投资机会

香港《稳定币条例》落地，以合规框架激活数字金融创新

稳定币是一种旨在与某些资产（通常是货币）维持相对稳定价值的虚拟资产。其主要特征在于通过足额资产储备或算法机制，将币值保持与目标资产（如美元、人民币、黄金等）基本挂钩，从而减小价格波动。稳定币既具备数字货币的便捷性和透明度，又具备法币资产的稳定性，被广泛用作区块链生态中的“链上现金”。

中心化模式的美元稳定币（USDT、USDC）为主导地位。按照维持价格稳定的不同机制划分，稳定币的主要类型包括法币支持型（以法币为抵押，如 USDC、USDT、TUSD、GUSD）、加密货币支持型、商品支持型和算法型等。USDT、USDC 作为美元稳定币的代表，其监管模式为香港法币稳定币的合规化提供了参考范本，而香港条例进一步强化了资本充足性（最低 2500 万港元）与储备隔离要求（定期审计披露）。香港通过立法明确稳定币的定义与类型，为沙盒参与者的合规发行提供了制度基础，也为跨境支付、RWA 等场景的落地铺平了道路。

图35：各类稳定币占比



资料来源：CoinGecko，国信证券经济研究所整理

图36：USDT 与 USDC 对照表

对比特征	USDT	USDC
发行方	Tether	Circle、Coinbase 等
挂钩或货币	1：1挂钩美元	1：1挂钩美元
储备资产	现金及等价物	美元存款、短期美债
透明度	未公开审计报告	每月公布储备情况

资料来源：银银平台订阅号，国信证券经济研究所整理

香港于 2025 年 5 月 21 日正式通过《稳定币条例》，旨在通过统一发牌制度、资本与储备监管、跨境流通规范，解决稳定币行业长期存在的透明度不足、赎回风险等问题，同时为合规机构开辟清晰的入场路径，将香港打造成亚太地区稳定币创新与监管的标杆。《稳定币条例》不仅是监管框架，更是香港布局数字金融的“场景激活器”。

核心监管逻辑：1）统一发牌制度：由香港金融管理局统一发牌，要求发行人实缴资本≥2500 万港元，强化资本充足性；2）储备资产隔离：稳定币需 100% 锚定高流动性资产（如现金、国债），并通过独立审计机构月度披露，确保每一枚稳定币均有足额资产支撑；3）多法币兼容：允许 USDC、USDT 等非港元稳定币合规流通，同时支持未来港元稳定币（如圆币科技 HKDR）发行，形成“美元稳定币对接国际市场 + 港元稳定币连接内地”的双轨格局。

主要受益方：“沙盒”参与者、跨境支付服务提供商、RWA 项目方三种类型主体有望充分收益。

图37：《稳定币条例》收益类型、政策要点及代表企业

收益类型	政策要点	代表企业
沙盒持牌机构	金管局统一发牌，沙盒参与者或将优先获牌	京东币链、渣打银行
跨境支付服务提供商	开放非港元稳定币流通，支持秒级跨境结算	华峰超纤、青岛金王
RWA 项目方	稳定币为资产代币化提供计价与结算工具	协鑫能科、郎新集团

资料来源：香港金管局官网，国信证券经济研究所整理

图38：稳定币利好“沙盒”参与者、跨境支付服务提供商、RWA 项目方三种类型主体

资料来源：香港金管局官网，国信证券经济研究所整理

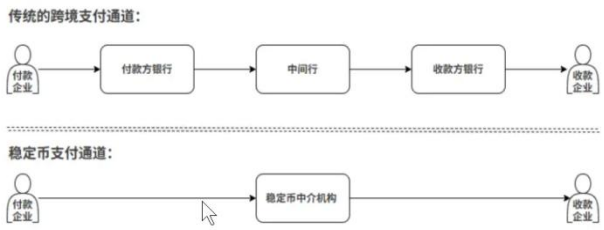
稳定币蓬勃发展，跨境支付+RWA 有望收益

跨境支付：支付效率与普惠性有望提升

稳定币在交易速度、成本以及体制方面显著优于传统跨境支付体系。稳定币基于区块链的点对点传输特性，既提升了全球资金流动效率、降低了交易成本，也拓展了金融服务的覆盖边界。正颠覆传统金融体系的“中介依赖”模式，而香港凭借区位优势与政策弹性，有望成为连接中国内地（CIPS）与全球稳定币网络的核心枢纽。

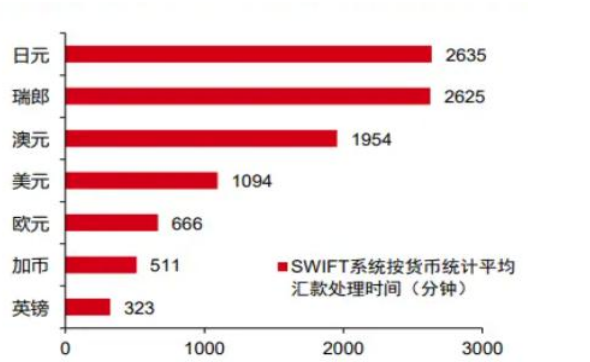
交易速度：绕开 SWIFT 体系的低效性。传统跨境汇款依赖 SWIFT 电文系统与代理银行网络，需经多级中转，平均耗时 1-3 天，手续费高达 5%-10%。稳定币基于区块链的点对点传输特性，实现了近乎实时的跨境转账（通常在几秒至几十秒内完成），且 7×24 小时运作，且单笔成本可压缩至 1 美元以内，若收发双方均接受稳定币结算，甚至可实现近乎零成本转移。

图39：稳定币支付缩短跨境交易结算时间



资料来源：Committee on Payments and Market Infrastructures、Swift GPI，国信证券经济研究所整理

图40：SWIFT 系统按货币统计平均汇款处理时间

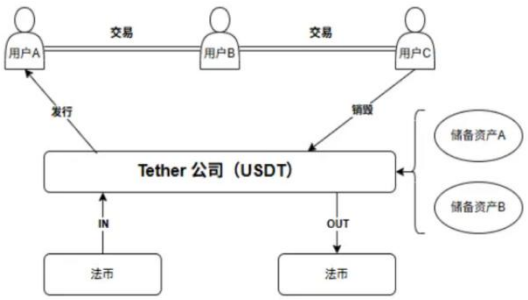


资料来源：Statrys，国信证券经济研究所整理

减少限制：突破法币兑换的中间壁垒。跨境交易涉及多币种兑换时，传统银行体系存在多重汇率加价（如美元→欧元→日元需两次兑换）。稳定币作为锚定单一法币的数字等价物（如 USDT 锚定美元），可直接作为 “中间货币” 完成多边兑换，减少汇率损失。

普惠性：重构全球金融包容性格局。2022 年，全球范围内仍有大约 17 亿成年人没有银行账户，但可通过智能手机使用稳定币钱包。2023 年，非洲数字支付的注册账户数达到了 8.56 亿个，占全球注册账户的五成，并贡献了全球注册账户总增长的 70%以上。这种 “去中介化” 模式使跨境支付从 “精英服务” 变为 “普惠工具”。

图41：USDT 发行和流通过程



资料来源：《Tether 白皮书：一种利用比特币区块链交易的法币代币》，国信证券经济研究所整理

图42：稳定币全年每天 24 小时不间断交易

资料来源：VISA，国信证券经济研究所整理

RWA：赋能 RWA 资产链上化与全球流通

我国 RWA 市场正处于快速发展期，呈现出多元且创新的态势。稳定币通过价值锚定、效率革命、合规护航、流动性激活，破解 RWA 项目的价格波动、跨境摩擦、信任缺失与门槛高等痛点，有望使协鑫光伏、元隆文创等案例实现收益率提升、成本下降、规模扩张、全球化渗透，为万亿级 RWA 市场奠定“技术 + 合规 + 场景”的落地基础，推动传统资产数字化转型效率提升，重塑全球资产配置规则，

成为连接实体与数字金融的核心枢纽。

图43：RWA 项目实践

	项目内容
郎新集团	郎新科技将旗下部分充电桩作为RWA锚定资产，借助蚂蚁链的区块链与IoT技术，使每个数字资产代表对应充电桩的部分收益权。
协鑫能科	协鑫能科以湖南、湖北82MW户用光伏电站为底层资产，发行RWA代币募资2亿元，为当时香港最大规模RWA项目，标志着国内首单光伏RWA的诞生。
元隆雅图	与香港胜利证券携手开拓IP文创周边RWA全球市场。
大连市小平岛	投资20亿元打造小平岛5C自驾车营地，成为国内首个RWA数字岛屿标杆项目。
巡鹰集团	国首个部署在区块链公链上的RWA项目。双方将安徽巡鹰新能源集团运营的约4000个电瓶车换电柜转化为数字金融产品，面向私募市场发行。
中国经济建设投资集团	首期融资3000万元，锚定“点点玉脉”平台和田玉手镯作为RWA资产。该平台完整整合玉石行业产业链，2025年5月，玉石RWA数字资产在青岛文化产权交易所上线，通过“海南数据跨境+香港资本通道”的创新模式实现全球化发行。

资料来源：公司公告，国信证券经济研究所整理

投资建议:看好 AI ASIC 及稳定币

全球 AI ASIC 快速发展，谷歌、亚马逊、Meta 等公司纷纷加快 ASIC 芯片的自研和测试；国内受 AI 芯片禁令影响，英伟达先进 AI 芯片对华出口受限，互联网大厂可能转向 AI ASIC 芯片，服务器厂商有望充分受益；同时，国产算力芯片 25 年有望快速放量，建议关注海光信息等。同时，香港《稳定币条例》落地，稳定币有望提升跨境支付效率，建议关注新大陆等公司。

风险提示

互联网大厂 AI ASIC 研发进展不及预期。

云厂商资本开支投入不及预期。

稳定币发行进展不及预期。

免责声明

分析师声明

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后 6 到 12 个月内的相对市场表现，也即报告发布日后的 6 到 12 个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A 股市场以沪深 300 指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数（HSI.HI）作为基准；美国市场以标普 500 指数（SPX.GI）或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数 10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数 10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数 10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数 10%以上

重要声明

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券经济研究所

深圳
深圳市福田区福华一路 125 号国信金融大厦 36 层
邮编：518046　总机：0755-82130833

上海
上海浦东民生路 1199 弄证大五道口广场 1 号楼 12 层
邮编：200135

北京
北京西城区金融大街兴盛街 6 号国信证券 9 层
邮编：100032