

# GPT 5后NLP大模型逐步走向收敛，ASIC将大有可为

## 行业研究·深度报告

### 投资评级：超配（维持评级）

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

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

证券分析师：朱松  
zhusong@guosen.com.cn  
S0980520070001

证券分析师：张伦可  
zhanglunke@guosen.com.cn  
S0980521120004

- **GPT 5后NLP大模型逐步走向收敛，ASIC将大有可为。** 理论来说，当大模型的算力、数据量以及参数量同时增大的时候，整个模型的效果会越来越好。GPT-3使用的数据量为40T，预测GPT-4的数据量需求为400T，但未来数据量将不会无限增大，所以当数据量不能增长的时候，光靠算力和模型参数量增长，并不能增强模型的效果。换言之，受到数据量的限制，NLP大语言模型的迭代或将进入停滞期。在模型迭代停滞期，若应用爆发，则ASIC芯片的优势就显现出来。ASIC针对特定需求开发，可根据需求对性能和功耗进行定向优化，其专用的芯片架构与高复杂度的算法相匹配，量产后在性能、功耗、成本方面均具有较大优势，ASIC芯片将大有可为。
- **头部厂商纷纷切入AI ASIC领域，技术路径不同。** 谷歌15年发布第一代TPU（ASIC）产品，TPU产品持续迭代升级；英特尔19年收购人工智能芯片公司Habana Labs，22年发布AI ASIC芯片Gaudi 2，性能表现出色；IBM研究院22年底发布AI ASIC芯片AIU，有望23年上市；三星第一代AI ASIC芯片Warboy NPU芯片已于近日量产。头部厂商纷纷切入 AI ASIC领域，看好ASIC在人工智能领域的长期成长性。
  - **谷歌：**谷歌为AI ASIC芯片的先驱，于15年发布第一代TPU（ASIC）产品，大幅提升AI推理的性能；17年发布TPU v2，在芯片设计层面，进行大规模架构更新，使其同时具备AI推理和AI训练的能力；谷歌TPU产品持续迭代升级，21年发布TPU v4，采用7nm工艺，峰值算力达275 TFLOPS，性能表现全球领先。
  - **英特尔：**19年底收购以色列人工智能芯片公司Habana Labs，22年发布Gaudi 2 ASIC芯片。从架构来看，Gaudi架构拥有双计算引擎（MME和TPC），可以实现MME和TPC并行计算，大幅提升计算效率；同时，其将RDMA技术应用于芯片互联，大幅提升AI集群的并行处理能力；从性能来看，Gaudi 2在ResNET-50、BERT、BERT Phase-1、BERT Phase-2模型的训练吞吐量优于英伟达A100，性能表现优异。
- **投资建议：建议关注海内外AI ASIC龙头。海外AI ASIC公司建议关注：英特尔、IBM等。**
- **风险提示：**宏观经济波动；下游需求不及预期；AI伦理风险；模型仍处于快速迭代期，ASIC芯片发展不及预期。

## 一、GPT 5后NLP大模型逐步走向收敛，ASIC将大有可为

- 1.1 ASIC具有性能高、体积小、功率低等特点
- 1.2 GPT 5后NLP大模型逐步走向收敛，ASIC将大有可为
- 1.3 ASIC芯片有望在推理领域率先出现爆品
- 1.4 全球AI芯片市场快速增长，ASIC占比将大幅提升

## 二、头部厂商纷纷切入AI ASIC领域，技术路径不同

- 2.1 谷歌——全球AI ASIC先驱，TPU产品持续迭代
- 2.2 英特尔——收购Habana Lab，Gaudi 2性能表现出色

## 三、投资建议：建议关注海内外AI ASIC龙头

## 1、GPT 5后NLP大模型逐步走向收敛，ASIC将大有可为

# ASIC具有性能高、体积小、功率低等特点

- ASIC具有性能高、体积小、功率低等特点。** 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。

表1：ASIC具有性能高、体积小、功率低等特点

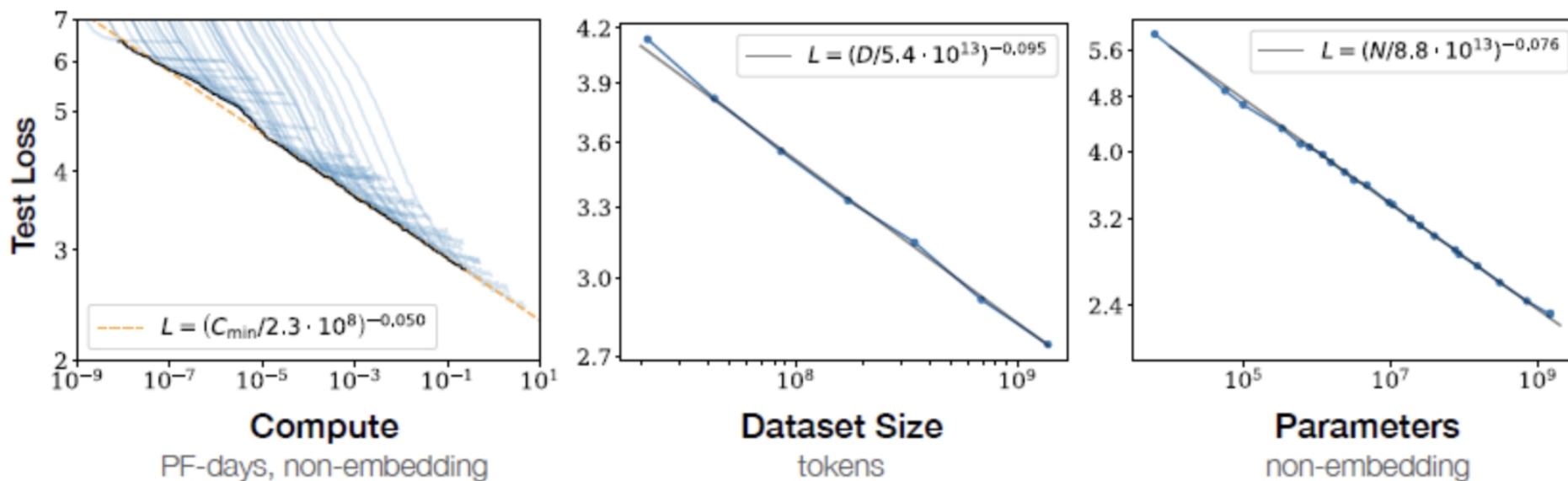
| 类别   | CPU                                                                                 | GPU                                                                                   | FPGA                                                                                  | ASIC                                                                                  |
|------|-------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| 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<br>✓ 通用性强<br>✓ 复杂指令和任务<br>✓ 系统管理                                              | ✓ 大量并行核<br>✓ AI处理出色表现                                                                 | ✓ 可配置的逻辑门<br>✓ 灵活性<br>✓ 可重新编程性                                                        | ✓ 可用库设计的定制化逻辑<br>✓ 更快的处理速度<br>✓ 体积小                                                   |
| 劣势   | ✓ 核数少<br>✓ 时延严重<br>✓ 效率低                                                            | ✓ 功耗高<br>✓ 体积大                                                                        | ✓ 编程复杂                                                                                | ✓ 固定的功能<br>✓ 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

# GPT 5后NLP大模型逐步走向收敛，ASIC将大有可为

- **NLP模型的迭代升级有可能进入停滞期。** 理论来说，当大模型的算力、数据量以及参数量同时增大的时候，整个模型的效果会越来越好。GPT-3使用的数据量为40T，预测GPT-4的数据量需求为400T，但未来数据量将不会无限增大，所以当数据量不能增长的时候，光靠算力和模型参数量增长，并不能增强模型的效果。换言之，受到数据量的限制，NLP大语言模型的迭代或将进入停滞期。
- **模型迭代停滞期，ASIC将大有可为：** 1) 模型快速迭代期：ASIC芯片的设计和制造需要大量资金、较长的时间周期和工程周期，且一旦定制无法再次进行写操作，所以在模型快速迭代期，ASIC不是最优解；2) 模型迭代停滞期：ASIC针对特定需求开发，可根据需求对性能和功耗进行定向优化，其专用的芯片架构与高复杂度的算法相匹配，量产后在性能、功耗、成本方面均具有较大优势，ASIC芯片将大有可为。

图1：算力、数据、参数量三者均变大的时候模型效果才会越来越好



# ASIC芯片有望在推理领域率先出现爆品

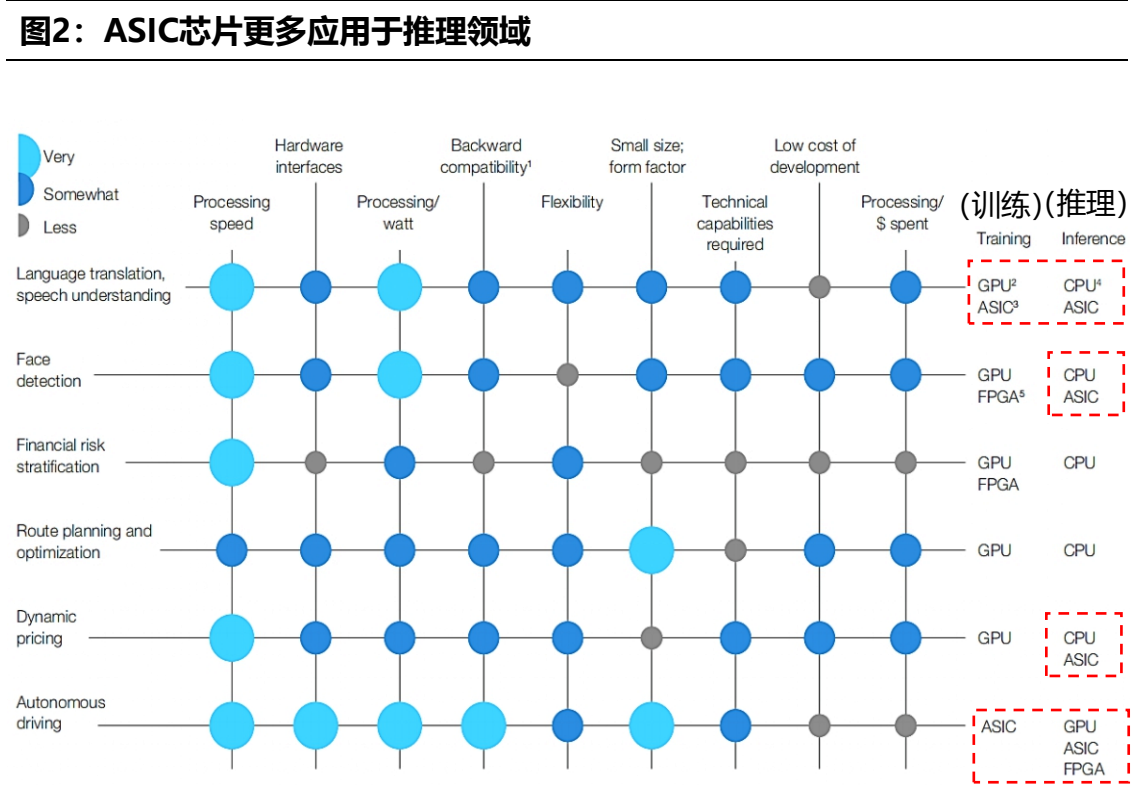
- 推理芯片更加关注功耗、时延、成本，且对精度要求相对较低。**深度学习由训练和推理两个任务组成，因为AI芯片的主要功能就是训练和推理。1) 训练：对大量数据在平台上进行学习，并形成具备特定功能的神经网络模型。由于训练所需数据量巨大、算法复杂度高，对AI芯片有高算力、高容量、高精度和通用性要求。2) 推理：指基于已经训练好的模型，针对输入数据计算结果，更加关注芯片功耗、时延、成本等因素，且对精度的要求相对较低。
- ASIC芯片在推理领域具有明显优势，有望在该领域率先出现爆品。**根据CSET数据，ASIC芯片在推理领域优势明显，其效率和速度约为CPU的100-1000倍，相较于GPU和FPGA具备显著竞争力。尽管ASIC芯片同样可以应用于训练领域（例如TPU v2、v3、v4），但我们认为其将在推理领域率先出现爆品。

| 表2：推理芯片算力、功耗较低 |                                |                              |
|----------------|--------------------------------|------------------------------|
| 芯片             | 昇腾Ascend910                    | 昇腾Ascend310                  |
| 功能             | 训练                             | 推理                           |
| 工艺/nm          | 7                              | 12                           |
| 算力             | INT8 640TOPS<br>FP16 320TFLOPS | INT8 22TOPS<br>FP16 11TFLOPS |
| 功耗/W           | 310                            | 8                            |
| 内存             | HBM2E                          | 2*LPDDR4x                    |

资料来源：华为、国信证券经济研究所整理

| 表3：ASIC芯片在推理领域具有明显优势 |             |             |            |           |     |           |
|----------------------|-------------|-------------|------------|-----------|-----|-----------|
|                      | 训练          |             | 推理         |           | 通用性 | 推理<br>准确率 |
|                      | 销量          | 速度          | 效率         | 速度        |     |           |
| CPU                  |             | 1x baseline |            |           | 很高  | ~98-99.7% |
| GPU                  | ~10-100x    | ~10-1,000x  | ~1-10x     | ~1-100x   | 高   | ~98-99.7% |
| FPGA                 | -           | -           | ~10-100x   | ~10-100x  | 中   | ~95-99%   |
| ASIC                 | ~100-1,000x | ~10-1,000x  | ~100-1000x | ~10-1000x | 低   | ~90-98%   |

资料来源：CSET、国信证券经济研究所整理

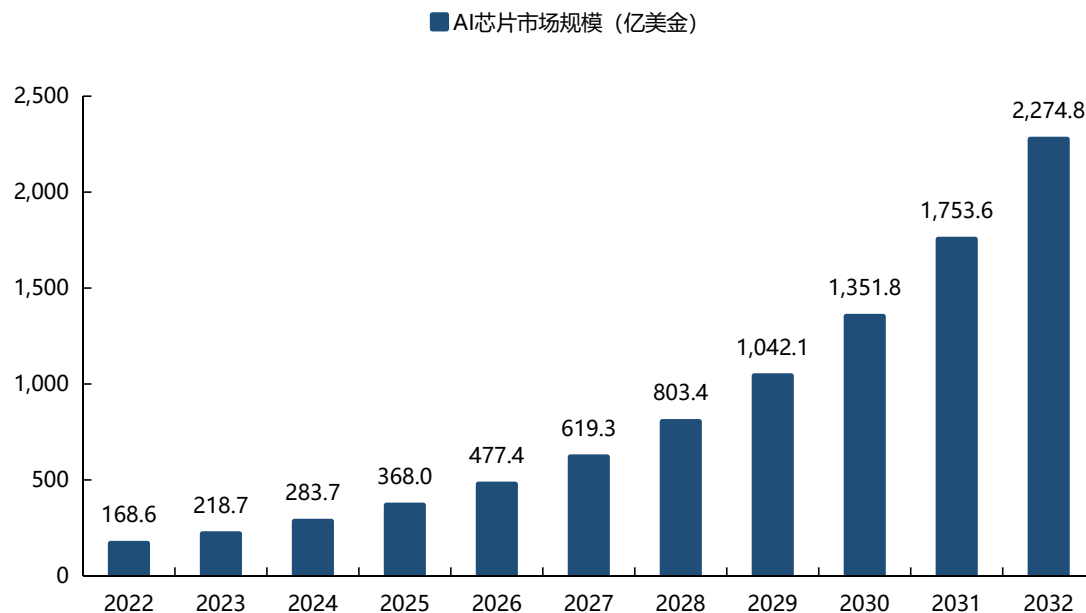


资料来源：McKinsey Analysis、国信证券经济研究所整理

# 全球AI芯片市场快速增长，ASIC占比将大幅提升

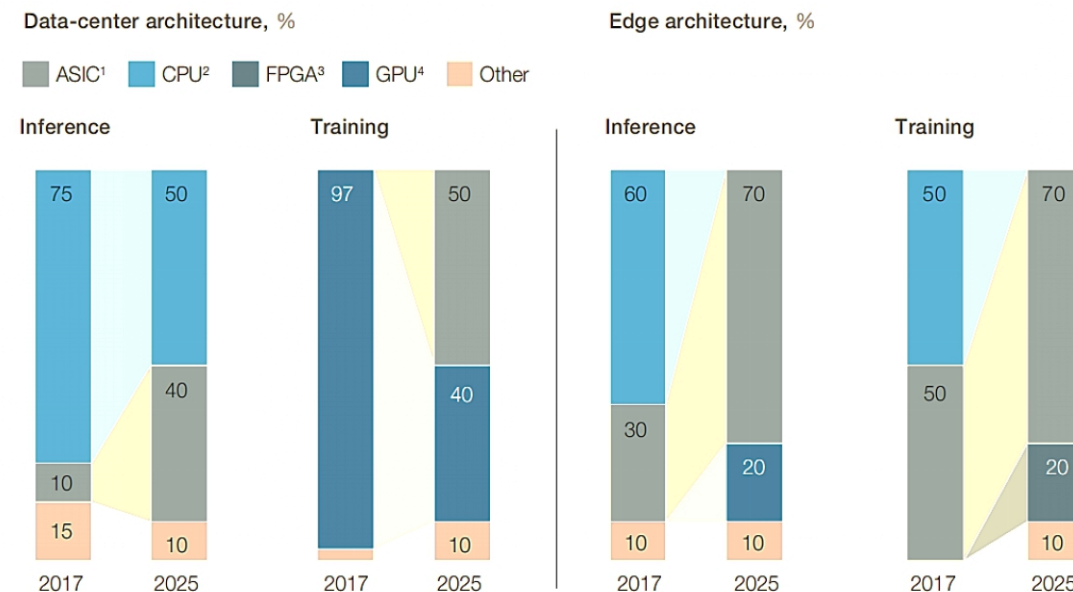
- **全球AI芯片市场快速增长，预计22-32年CAGR达29.72%。**根据Precedence Research数据，22年全球AI芯片市场规模约为168.6亿美金；随着专用AI向通用AI的发展，以及AI应用领域持续拓宽，算力需求爆发，拉动全球AI芯片市场快速增长，预计2032年将达到2,274.8亿美金，对应22-32年CAGR达29.72%。
- **预计ASIC在AI芯片的占比将大幅提升。**根据McKinsey Analysis数据，在数据中心侧，25年ASIC在推理/训练应用占比分别达到40%、50%；在边缘侧，25年ASIC在推理/训练应用占比分别达到70%、70%，ASIC在AI芯片的占比将大幅提升。

图3：22年全球AI芯片市场规模约为168.6亿美金，22-32年CAGR为29.72%



资料来源：Precedence Research、国信证券经济研究所整理

图4：预计ASIC在AI芯片的占比将大幅提升



资料来源：McKinsey Analysis、国信证券经济研究所整理

## 2、头部厂商纷纷切入AI ASIC领域，技术路径不同

# 多种类AI芯片并存，头部厂商纷纷切入ASIC领域



- **多种类AI芯片并存，头部厂商纷纷切入ASIC领域。**英伟达延续GPU路线，22年发布H100芯片，目前广泛应用于云端训练和推理；AMD利用自身技术积累，将CPU和GPU集成在一起，推出Instinct MI300芯片，预计23年H2上市。头部厂商开始切入ASIC领域，Google为AI ASIC芯片的先驱，21年推出TPU v4，运算效能大幅提升；英特尔19年收购Habana Lab，22年推出Gaudi2 ASIC芯片；IBM、三星等头部厂商亦纷纷切入ASIC领域。

表4：头部厂商开始切入ASIC领域

| 序号 | 公司  | 典型AI芯片         | 发布年份        | 技术架构             | 功能任务      |
|----|-----|----------------|-------------|------------------|-----------|
| 1  | 英伟达 | H100           | 2022        | GPU              | 云端训练、云端推理 |
|    |     | A100           | 2020        | GPU              | 云端训练、云端推理 |
|    |     | V100           | 2017        | GPU              | 云端训练、云端推理 |
| 2  | 英特尔 | Gaudi2         | 2022        | ASIC             | 云端训练、云端推理 |
|    |     | Nervana NNP-T  | 2019        | NNP-T1000        | 云端训练      |
|    |     | Nervana NNP-I  | 2019        | NNP-I1000        | 云端推理      |
| 3  | IBM | AIU            | 2023        | ASIC             | 云端训练      |
|    |     | TrueNorth      | 2015        | 类脑芯片             | 边缘端推理     |
| 4  | 谷歌  | TPU v4         | 2021        | ASIC             | 云端训练、云端推理 |
|    |     | TPU v3         | 2018        | ASIC             | 云端训练、云端推理 |
|    |     | Edge TPU       | 2018        | ASIC             | 边缘端推理     |
| 5  | 苹果  | A14            | 2020        | ARM架构SoC         | 边缘端推理     |
| 6  | AMD | Instinct MI300 | 2023        | APU              | 云端训练      |
|    |     | EPYC2          | 2019        | Zen2架构           | 云端推理      |
| 7  | ARM | ARM Cortex-M55 | 2020        | ARM Helium       | 边缘端推理     |
|    |     | ARM Ethos-U55  | 2020        | ASIC (Micro NPU) | 边缘端推理     |
| 8  | 高通  | 骁龙888          | 2020        | ARM架构SoC         | 边缘端推理     |
|    |     | Cloud AI 100   | 2020        | ASIC             | 云端推理      |
| 9  | 三星  | Warboy         | 2023 (开始量产) | ASIC (NPU)       | 云端训练、云端推理 |
|    |     | Exynos2100     | 2021        | ARM架构SoC         | 边缘端推理     |

资料来源：各企业官网、商惠敏-《人工智能芯片产业技术发展研究》-全球科技经济瞭望（2021年）-P27、国信证券经济研究所整理

# 谷歌：全球AI ASIC先驱，TPU产品持续迭代

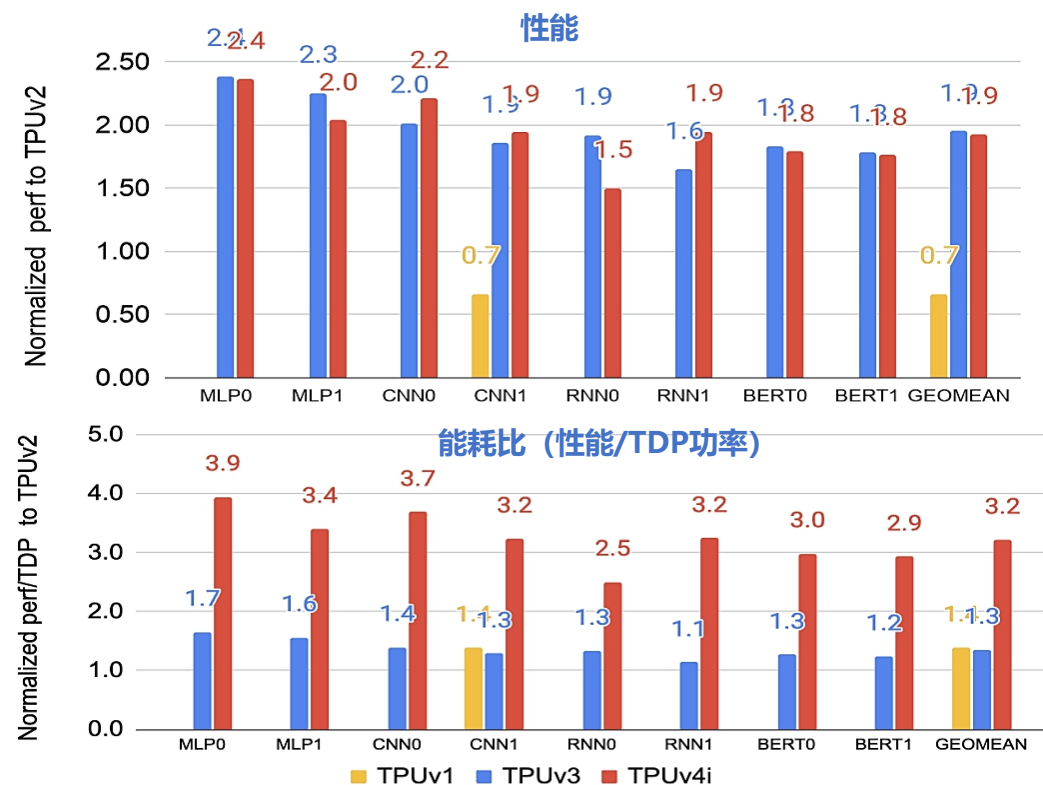
- 谷歌为全球AI ASIC先驱，TPU产品持续迭代。谷歌2015年发布TPU v1，与使用通用CPU和GPU的神经网络计算相比，TPU v1带来了15~30倍的性能提升和30~80倍的能效提升，其以较低成本支持谷歌的很多服务，仅可用于推理；17年发布TPU v2，用于加速大量的机器学习和人工智能工作负载，包括训练和推理；18年发布TPU v3，算力和功率大幅增长，其采用了当时最新的液冷技术；20年和21年分别发布TPU v4i和v4，应用7nm工艺，晶体管数大幅提升，算力提升，功耗下降。

表5：Google TPU产品持续迭代

| 芯片          | TPU v1              | TPU v2              | TPU v3              | TPU v4i             | TPU v4              |
|-------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| 发布时间        | 2015                | 2017                | 2018                | 2020                | 2021                |
| 推理          | √                   | √                   | √                   | √                   | √                   |
| 训练          | ×                   | √                   | √                   | ×                   | √                   |
| 峰值算力（单片）    | 92 TFLOPS           | 46 TFLOPS           | 123 TFLOPS          | 138 TFLOPS          | 275 TFLOPS          |
| 工艺          | 28 nm               | 16 nm               | 16 nm               | 7 nm                | 7 nm                |
| 裸片尺寸        | 330 mm <sup>2</sup> | 625 mm <sup>2</sup> | 700 mm <sup>2</sup> | 400 mm <sup>2</sup> | 780 mm <sup>2</sup> |
| 晶体管数        | 3 B                 | 9 B                 | 10 B                | 16 B                | 31 B                |
| 时钟速度        | 700 MHz             | 700 MHz             | 940 MHz             | 1,050 MHz           | 1,050 MHz           |
| TensorCores | 1颗                  | 2颗                  | 2颗                  | 1颗                  | 2颗                  |
| HBM存储带宽     | 300 Gb/sec          | 700 Gb/sec          | 900 Gb/sec          | 300 Gb/sec          | 1,200 Gb/sec        |
| 矩阵输入        | INT8                | BF16                | BF16                | BF16/INT8           | BF16/INT8           |
| 热设计功率（单片）   | 75                  | 280                 | 450                 | 175                 | 300                 |

资料来源：Google、Norman PJ等-《Ten Lessons From Three Generations Shaped Google's TPUv4i: Industrial Product》-ISCA (2021) -P2、国信证券经济研究所整理

图5：TPU v4i的性能和能耗比大幅提升（下图基准为TPU v2）

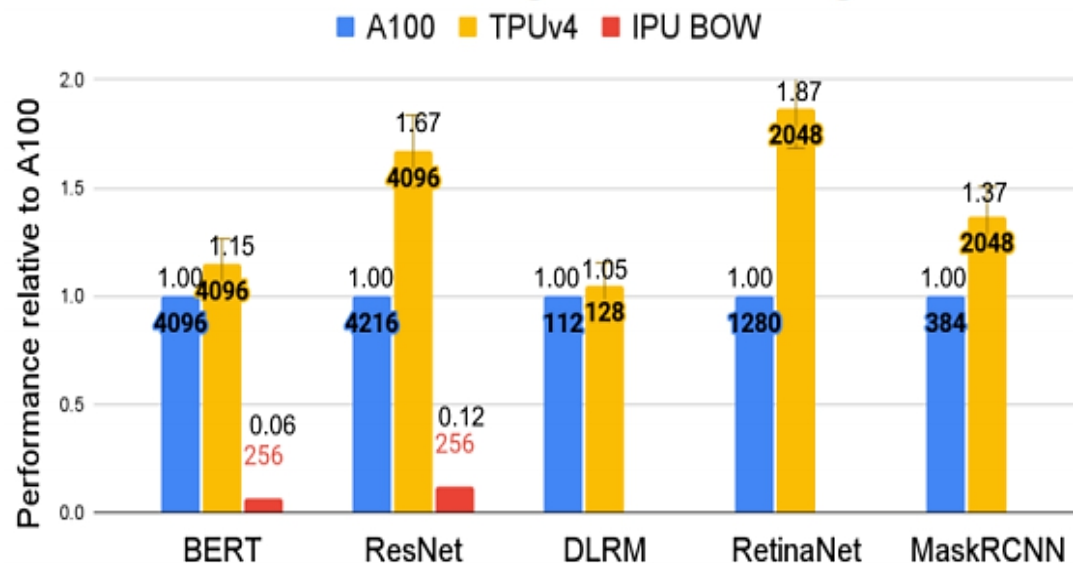


资料来源：Norman PJ等-《Ten Lessons From Three Generations Shaped Google's TPUv4i: Industrial Product》-ISCA (2021) -P9、国信证券经济研究所整理

# 谷歌：性能表现，A100<TPU v4<H100

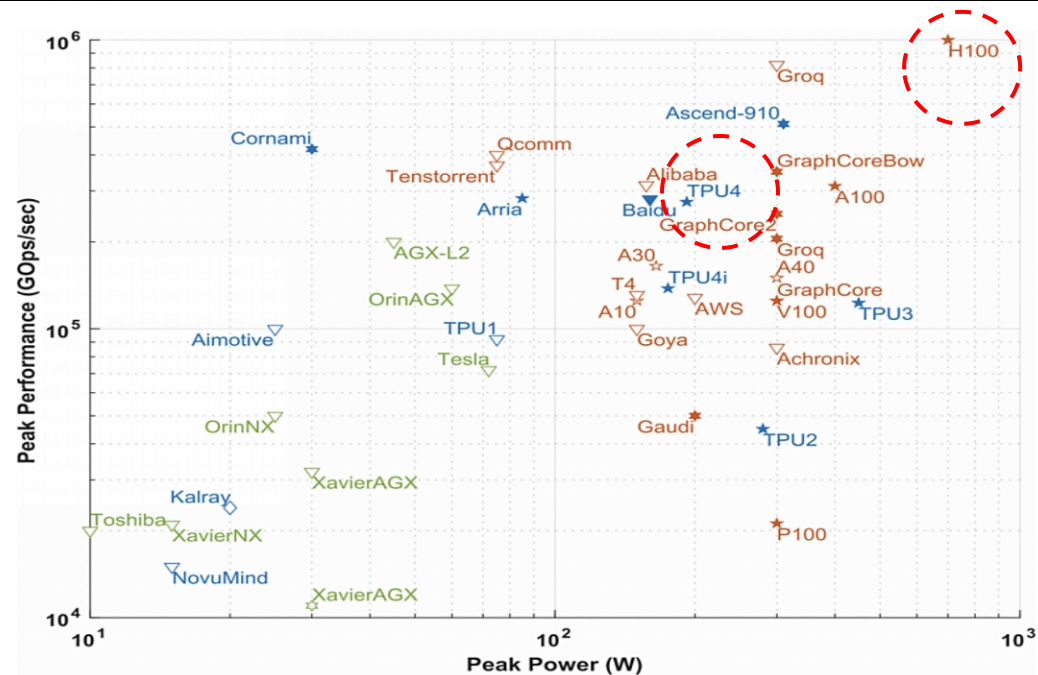
- **TUP v4性能表现优于英伟达A100。**根据《TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings》数据，TPU v4的性能表现在BERT、ResNet、DLRM、RetinaNet、MaskRCNN下分别为A100的1.15x、1.67x、1.05x、1.87x和1.37x，性能表现优于英伟达A100。
- **TUP v4性能表现略逊于H100，但功耗管理能力出色。**根据《AI and ML Accelerator Survey 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6：Google TPU产品持续迭代



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

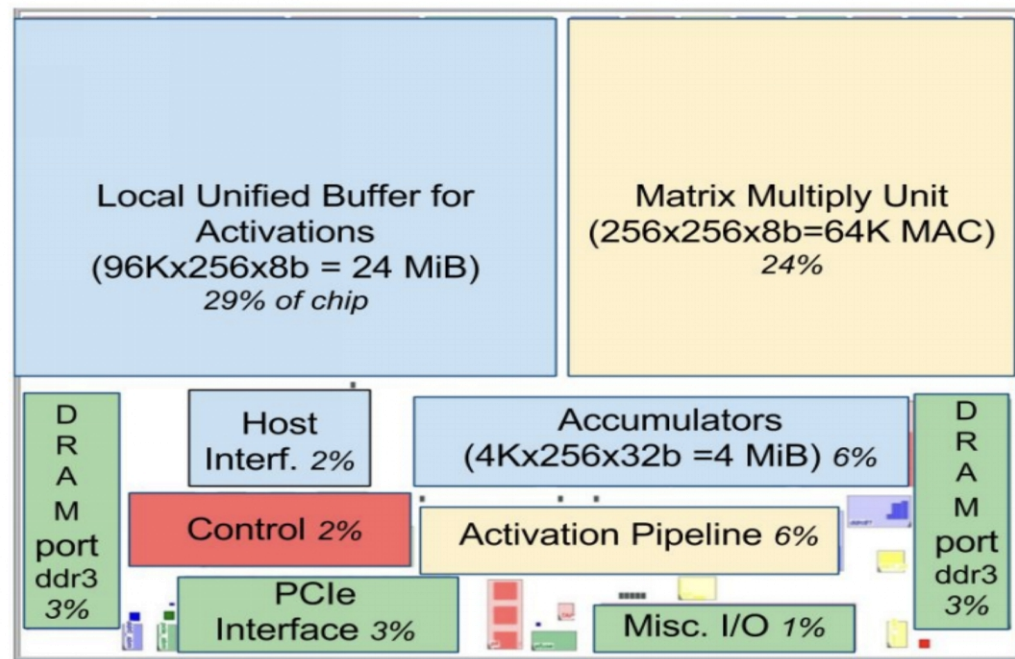
图7：TPU v4i的性能和能耗比大幅提升（下图基准为TPU v2）



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

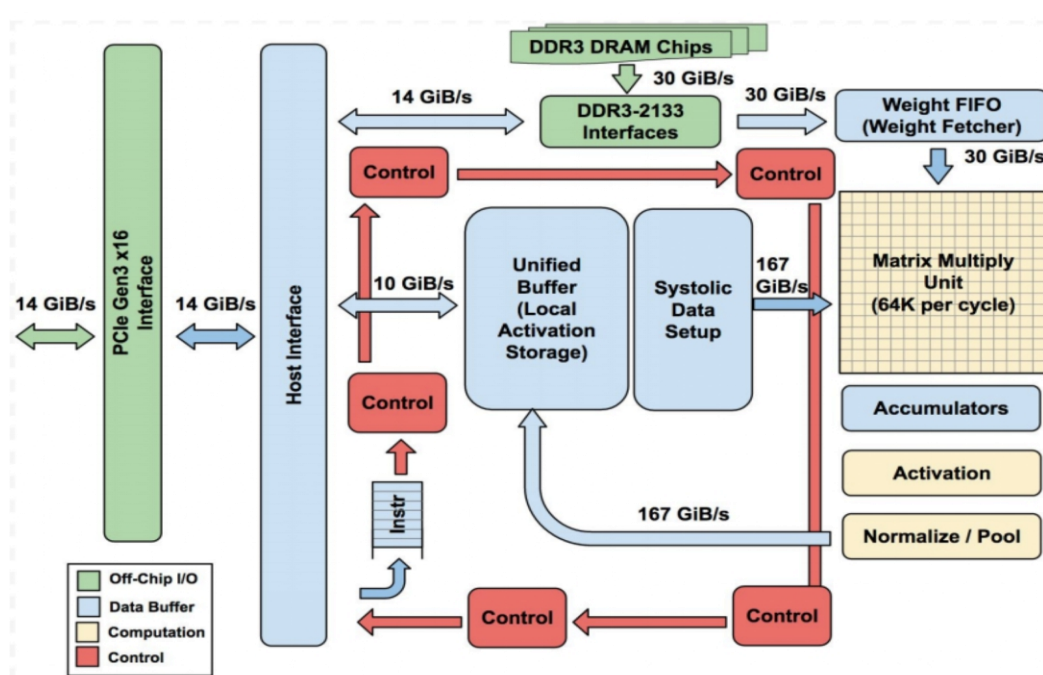
- **统一缓冲器 (Unified Buffer) 和矩阵乘法单元 (MMU) 占据53%的芯片总面积。** TPU v1主要包括统一缓冲器 (Unified Buffer)、矩阵乘法单元 (MMU)、累加器 (Accumulators)、激活流水线电路 (Activation Pipeline)、DDAM等，其中统一缓冲器和矩阵乘法单元面积占比最高，合计达53%。
- **TPU v1工作流程：** 1) 芯片启动，缓冲区和DDR3为空；2) 用户加载TPU编译的模型，将权重放入DDR3内存；3) 主机用输入值填充激活缓冲区；4) 发送控制信号将一层权重加载到矩阵乘法单元；5) 主机触发执行，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8：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9：TPU v1 Block Diagram

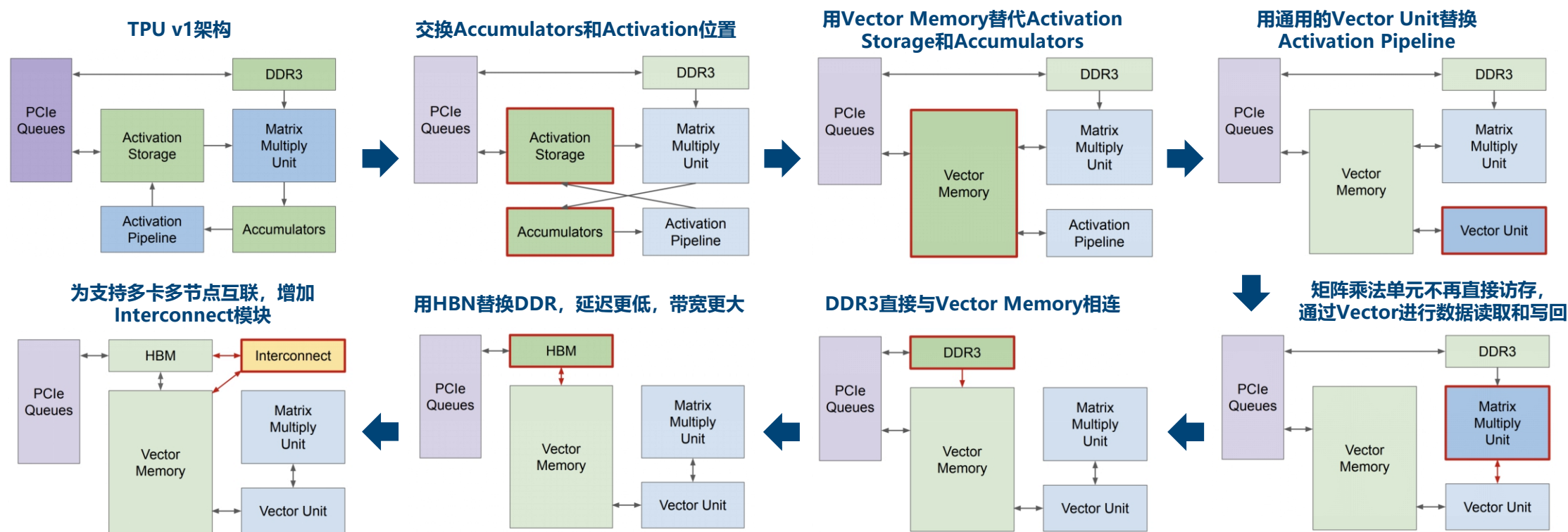


资料来源：Norman P.J等-《In-Datacenter Performance Analysis of a Tensor Processing Unit》-ISCA (2017) -P3、国信证券经济研究所整理

- 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10：TPU v1向TPU v2架构演进

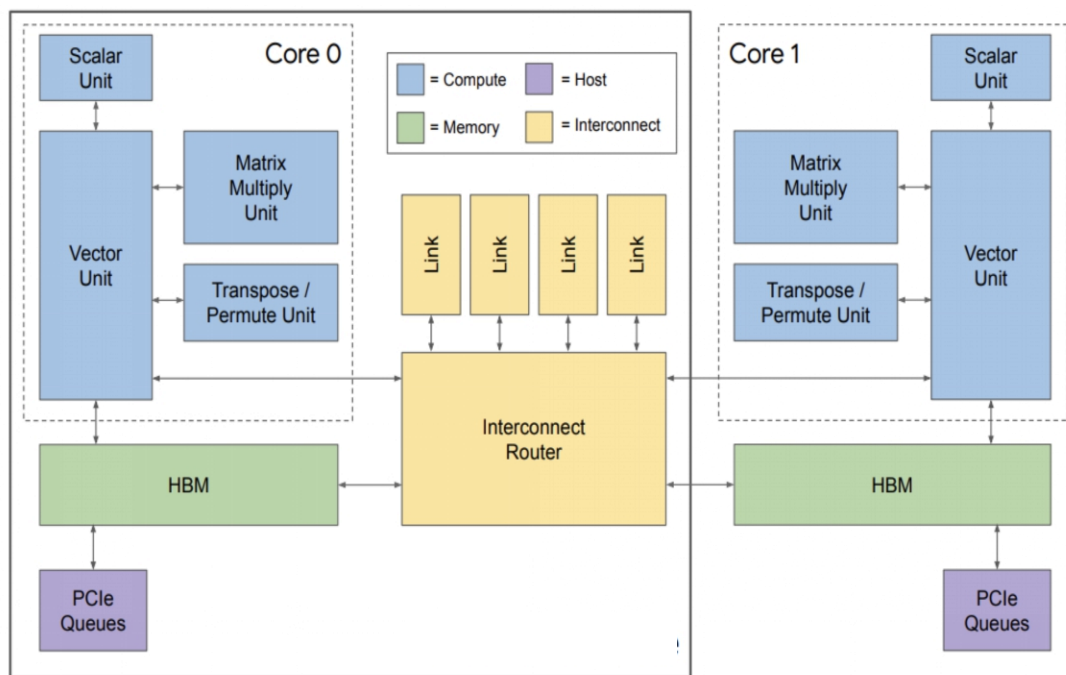


资料来源：Norrie T等-《The Design Process for Google's Training Chips: TPuv2 and TPuv3》-IEEE (2020) -P3、国信证券经济研究所整理

# 谷歌：TPU v2内核数增加至2个

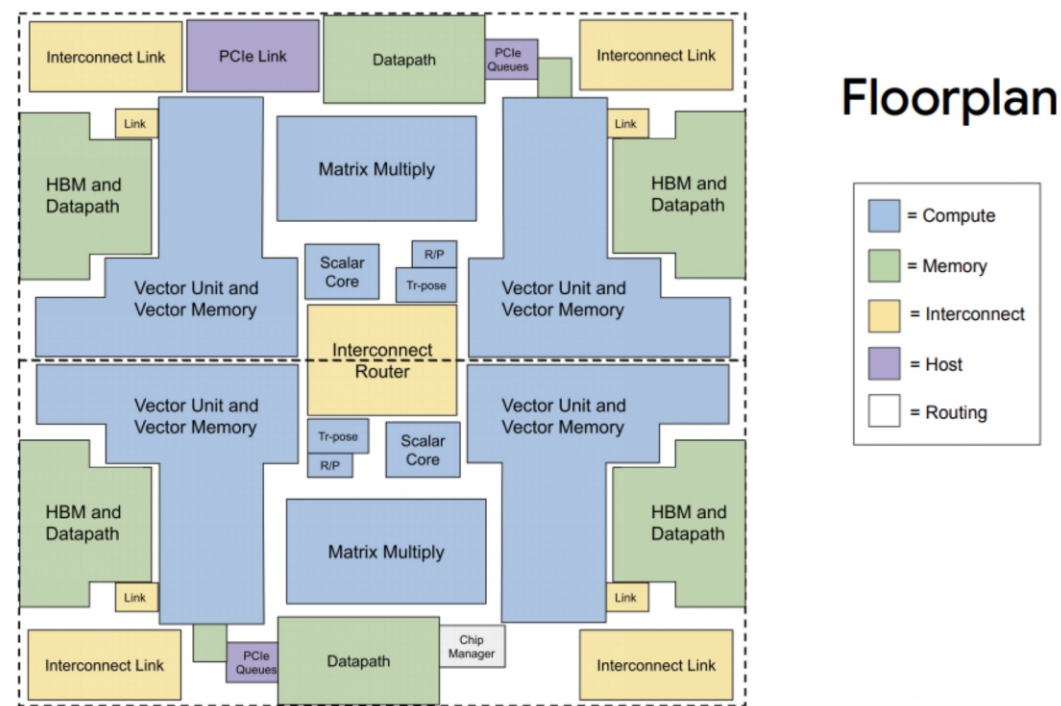
- **TPU v2内核数增加值2个。** TPU v1仅有1个Tensor Core，导致管道更为冗长。TPU v2的内核数增加为2个，对编译器也更为友好。
- **MXU利用率提升。** TPU v1的MXU包含256\*256个乘积累加运算器，由于部分卷积计算规模小于256\*256，导致单个大核的利用率相对较低；而TPU v2的单核MXU包含128\*128个乘积累加运算器，在一定程度上，提升了MXU利用率。

图11：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12：TPU v2 Floorplan

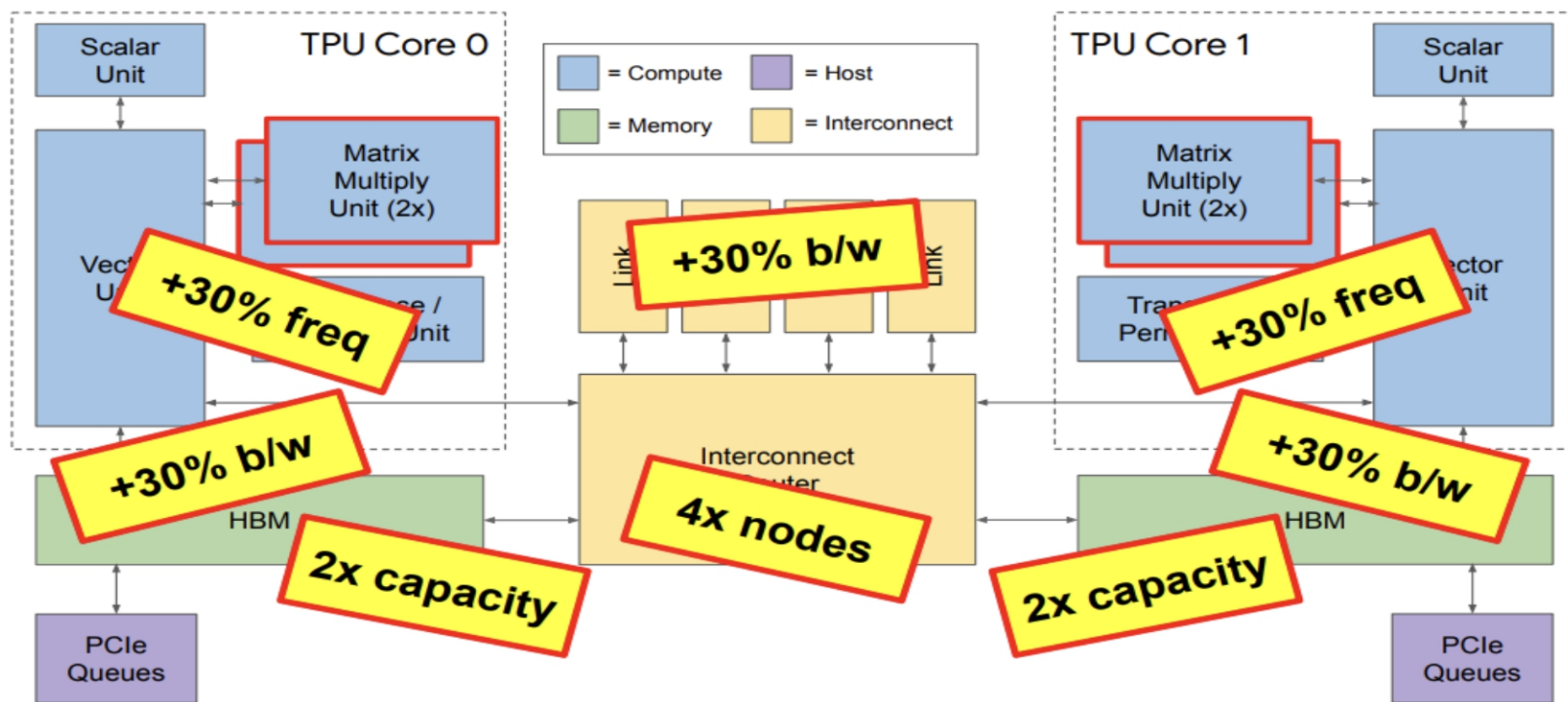


资料来源：Norrie T等-《The Design Process for Google' s Training Chips: TPUv2 and TPUv3》-IEEE (2020) -P7、国信证券经济研究所整理

# 谷歌：TPU v3延续v2架构，性能提升，TDP优化

- **谷歌TPU v3延续v2架构，性能提升。** 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。
- **采用液冷技术，TDP优化。** TPU v3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。

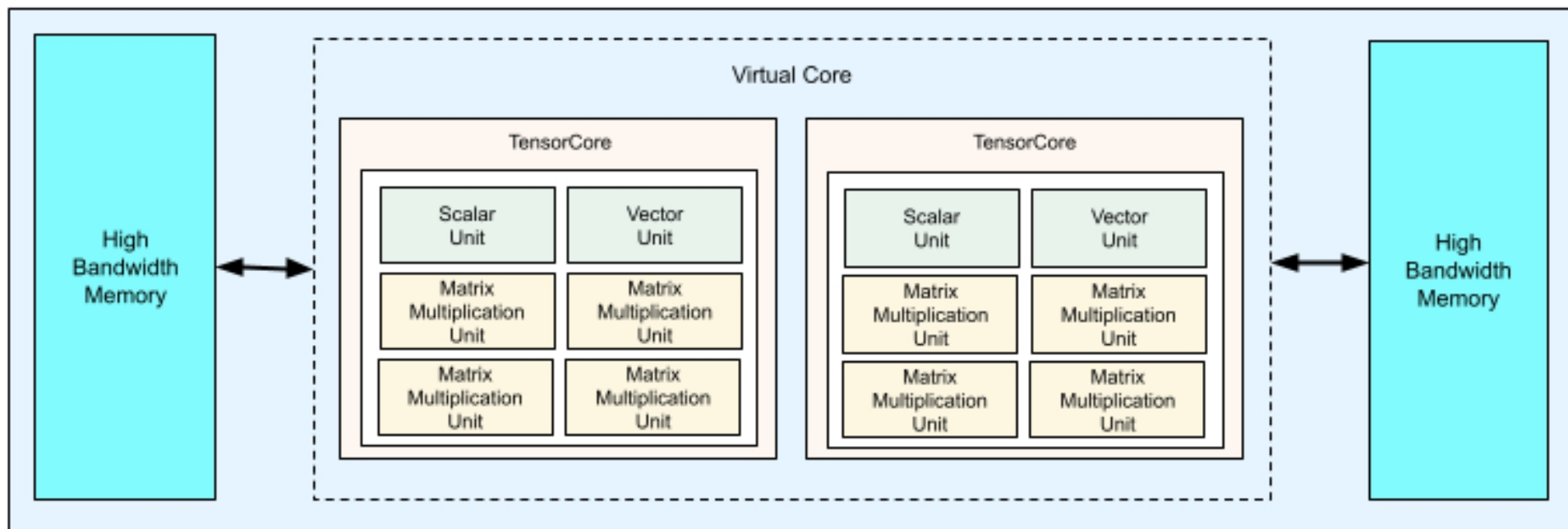
图13：TPU v3延续v2架构



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

- **MXU数量翻倍，峰值算力大幅提升。**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。

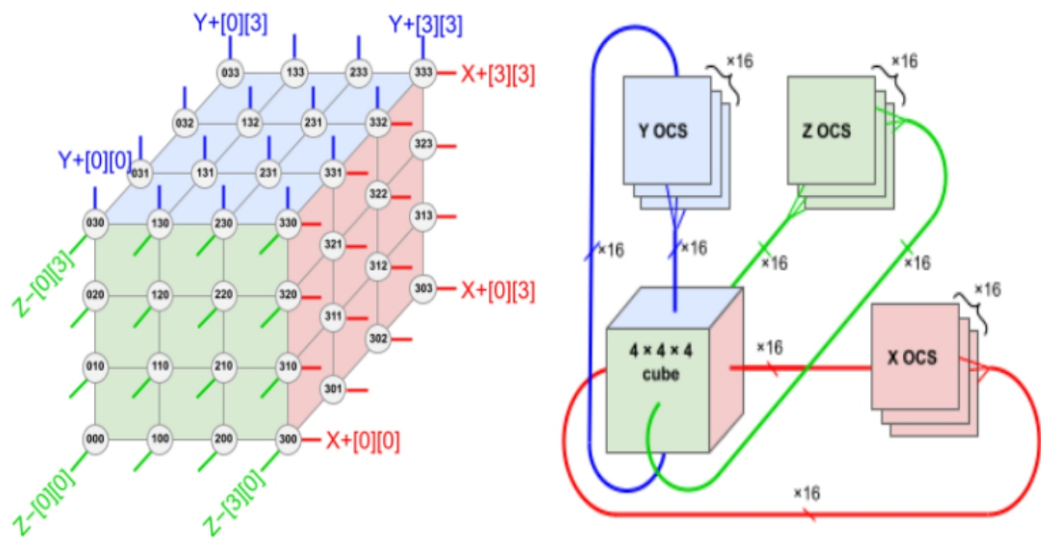
图14：TPU v4 MXU数量翻倍，峰值算力大幅提升



资料来源：Google Cloud、国信证券经济研究所整理

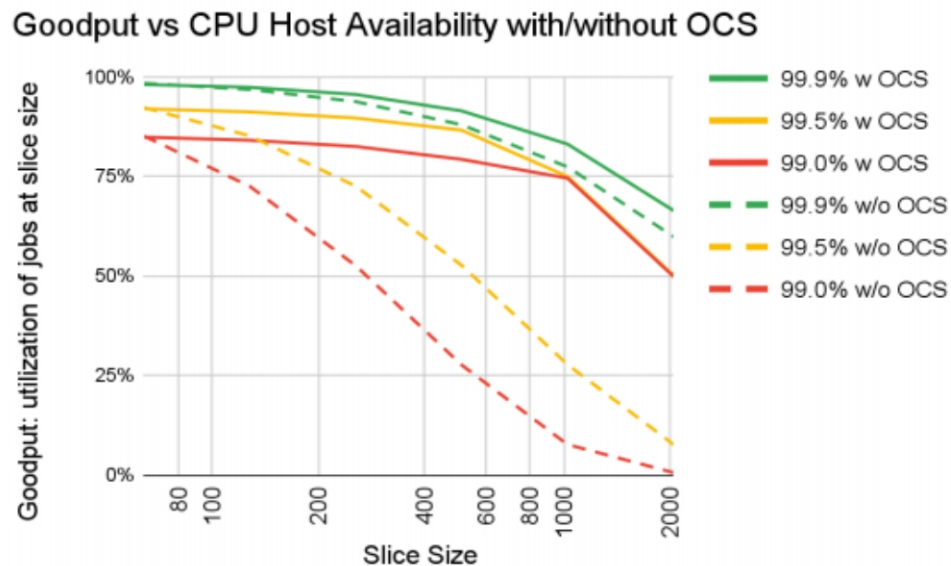
- **谷歌的超级计算机构想：**将4\*4\*4（64）个TPU v4芯片连接成1个立方体结构（Cube），再将4\*4\*4个立方体结构（Cube）连接成共有4096个TPU v4芯片的超级计算机，其中物理距离较近TPU v4芯片（即同一个Cube中的4\*4\*4个芯片）采用常规电互联方式，距离较远的TPU（例如Cube之间的互联）间用光互连。采用光互连技术可以有效避免“芯片等数据”的情形出现，进而提升计算效率。
- **可重配置光互连技术可以进一步提升计算性能。**谷歌TPU v4通过加入光路开关（OCS）的方式，可以根据具体模型数据流来调整TPU之间的互联拓扑，实现最优性能，根据《TPU 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。
- **可重配置光互连技术提升计算机的稳定性。**若计算机中部分芯片出现故障，可以通过该技术绕过故障芯片，进而不会影响整个系统的工作。

图15：谷歌超级计算机互联结构



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16：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

# 英特尔：Gaudi架构实现MME和TPC并行运算

- **英特尔收购Habana Lab。** Habana Labs成立于2016年，总部位于以色列，是一家为数据中心提供可编程深度学习加速器厂商，2019年发布第一代Gaudi。英特尔于2019年底收购Habana Lab，旨在加快其在人工智能芯片领域的发展，2022年发布Gaudi 2。
- **Gaudi架构实现MME和TPC并行运算。** Gaudi架构包含2个计算引擎，即矩阵乘法引擎（MME）和TPC（张量处理核心）；Gaudi架构使得MME和TPC计算时间重叠，进行并行运算，进而大幅提升计算效率。
- **Gaudi 2延续上一代架构，硬件配置大幅提升。** Gaudi 2架构基本与上一代相同，TPC数量从8个提升至24个，HBM数量从4个提升至6个（总内存从32GB提升至96GB），SRAM存储器提升一倍，RDMA从10个提升至24个，同时集成了多媒体处理引擎，硬件配置大幅提升。

图17：Gaudi架构实现MME和TPC并行运算，大大提升计算效率

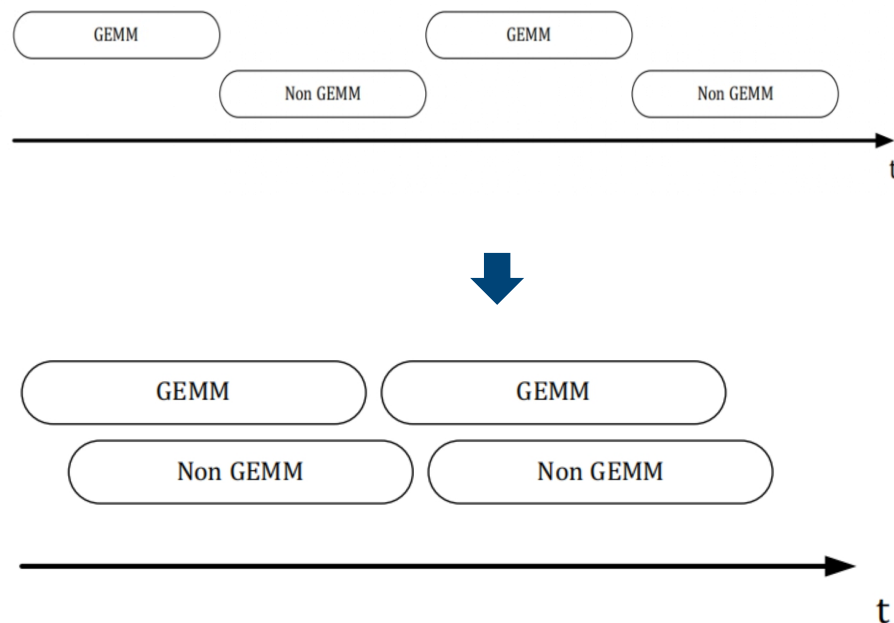
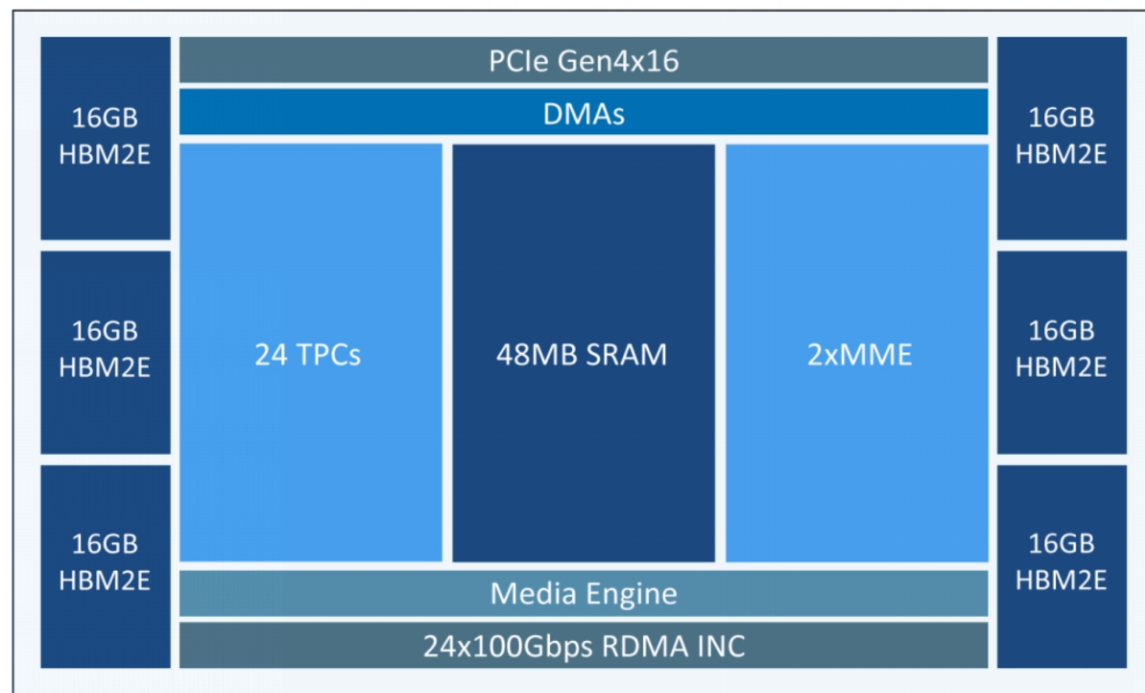
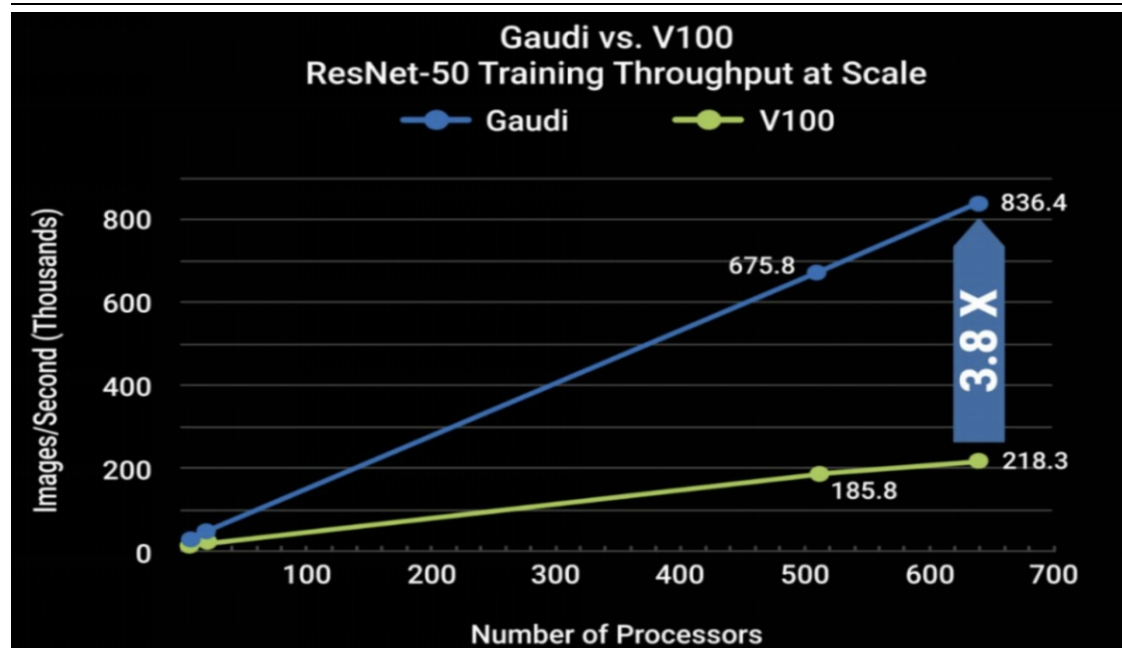


图18：Gaudi 2 Block Diagram



- **RDMA技术用于芯片互联，大幅提升并行处理能力。** RDMA是一种远端内存直接访问技术，具有高速、超低延迟和极低CPU使用率的特点。Gaudi将RDMA集成在芯片上，用于实现芯片间互联，大幅提升AI集群的并行处理能力；同时，Gaudi支持通用以太网协议，客户可以将Gaudi放入现有的数据中心，使用标准以太网构建AI集群。
- **Gaudi 2性能表现出色。** 根据《Habana Gaudi 2 White Paper》披露数据，Gaudi 2在ResNET-50、BERT、BERT Phase-1、BERT Phase-2模型的训练吞吐量分别为A100 (40GB, 7nm) 的2.0、2.4、2.1、3.3x，性能表现出色。

图19：RDMA大幅提升Gaudi并行处理能力



资料来源：Intel Habana、国信证券经济研究所整理

图20：Gaudi 2性能表现出色



资料来源：《Habana Gaudi 2 白皮书》、国信证券经济研究所整理

### 3、投资建议：建议关注海内外AI ASIC龙头

- **GPT 5后NLP大模型逐步走向收敛，ASIC迎来黄金发展期，建议关注海内外AI ASIC龙头。** GPT 5后NLP大模型受到数据量限制，或将逐步走向收敛，进入模型迭代停滞期，若应用爆发则ASIC的优势就显现出来了。ASIC针对特定需求开发，可根据需求对性能和功耗进行定向优化，其专用的芯片架构与高复杂度的算法相匹配，量产后在性能、功耗、成本方面均具有较大优势；ASIC是模型停滞期的最优解，将迎来黄金发展期，建议重点关注海内外AI ASIC龙头公司。
- **英特尔 (INTC.O)：** 英特尔19年底收购以色列人工智能芯片公司Habana Labs，22年发布Gaudi 2 ASIC芯片。从架构来看，Gaudi架构拥有双计算引擎（MME和TPC），可以实现MME和TPC并行计算，大幅提升计算效率；同时，其将RDMA技术应用于芯片互联，大幅提升AI集群的并行处理能力。从性能来看，Gaudi 2在ResNET-50、BERT、BERT Phase-1、BERT Phase-2模型的训练吞吐量优于英伟达A100，性能表现优异。
- **IBM (IBM.N)：** IBM研究院22年底发布AI ASIC芯片AIU，旨在支持多种格式并简化从图像识别到自然语言处理的人工智能工作流程。AIU包含32个基于5nm技术的处理核心以及230亿个晶体管，有望于23年上市。

第一，宏观经济波动。若宏观经济波动，产业变革及新技术的落地节奏或将受到影响，宏观经济波动还可能对 IT 投资产生负面影响，从而导致整体行业增长不及预期。

第二，下游需求不及预期。若下游数字化需求不及预期，相关的数字化投入增长或慢于预期，致使行业增长不及预期。

第三，AI伦理风险: AI可能会生产违反道德、常规、法律等内容。

第四，模型仍处于快速迭代期，ASIC芯片发展不及预期。ASIC芯片的设计和制造需要大量资金、较长的时间周期和工程周期，且一旦定制无法再次进行写操作，若模型仍处于快速迭代期，ASIC芯片发展不及预期。

| 国信证券投资评级 |    |                            |
|----------|----|----------------------------|
| 类别       | 级别 | 定义                         |
| 股票投资评级   | 买入 | 预计6个月内，股价表现优于市场指数20%以上     |
|          | 增持 | 预计6个月内，股价表现优于市场指数10%-20%之间 |
|          | 中性 | 预计6个月内，股价表现介于市场指数±10%之间    |
|          | 卖出 | 预计6个月内，股价表现弱于市场指数10%以上     |
| 行业投资评级   | 超配 | 预计6个月内，行业指数表现优于市场指数10%以上   |
|          | 中性 | 预计6个月内，行业指数表现介于市场指数±10%之间  |
|          | 低配 | 预计6个月内，行业指数表现弱于市场指数10%以上   |

## 分析师承诺

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。

## 重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

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  
GUOSEN SECURITIES

## 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楼

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