西南证券
SOUTHWEST SECURITIES

科技前瞻专题

AI ASIC：算力芯片的下一篇章

西南证券研究发展中心
海外研究团队
2024年12月

投资逻辑

- **ASIC 可以适应不同的业务场景和商业模式的需求，可以满足大型CSP客户的诸多需求**：1）内部工作负载的架构优化；2）更低的功耗，更低的成本；3）为AI工作负载定制的内存和I/O架构。随着AI应用的发展和生态逐步完善，AI算力集群特别是推理集群对加速计算芯片需求巨大，驱动ASIC快速成长。预计2028年数据中心 ASIC 市场规模将提升至429亿美元，CAGR为45.4%。

- **ASIC针对特定算法和应用进行优化设计，在特定任务上的计算能力强大，通常具有较高的能效比。**目前ASIC以推理场景应用为主，并开始切入到部分训练环节。对照北美四大CSP的自研产品路线：Google的TPU出货目前以v5产品为主，2025年将量产TPU v6；亚马逊的ASIC产品包括Trainium和Inferentia，分别用于训练和推理环节；微软和Meta也推出了各自的ASIC产品Maia 100和MTIA。由于大型CSP的业务模型、应用场景等多通过自身云来承载，每个云承载了独特的应用和商业模型，包括内部应用（比如搜索引擎、社交媒体等）、SaaS服务（比如AI聊天机器人、Copilot等）、IaaS服务等，自研ASIC可适应自身不同的业务场景和商业模式的需求。

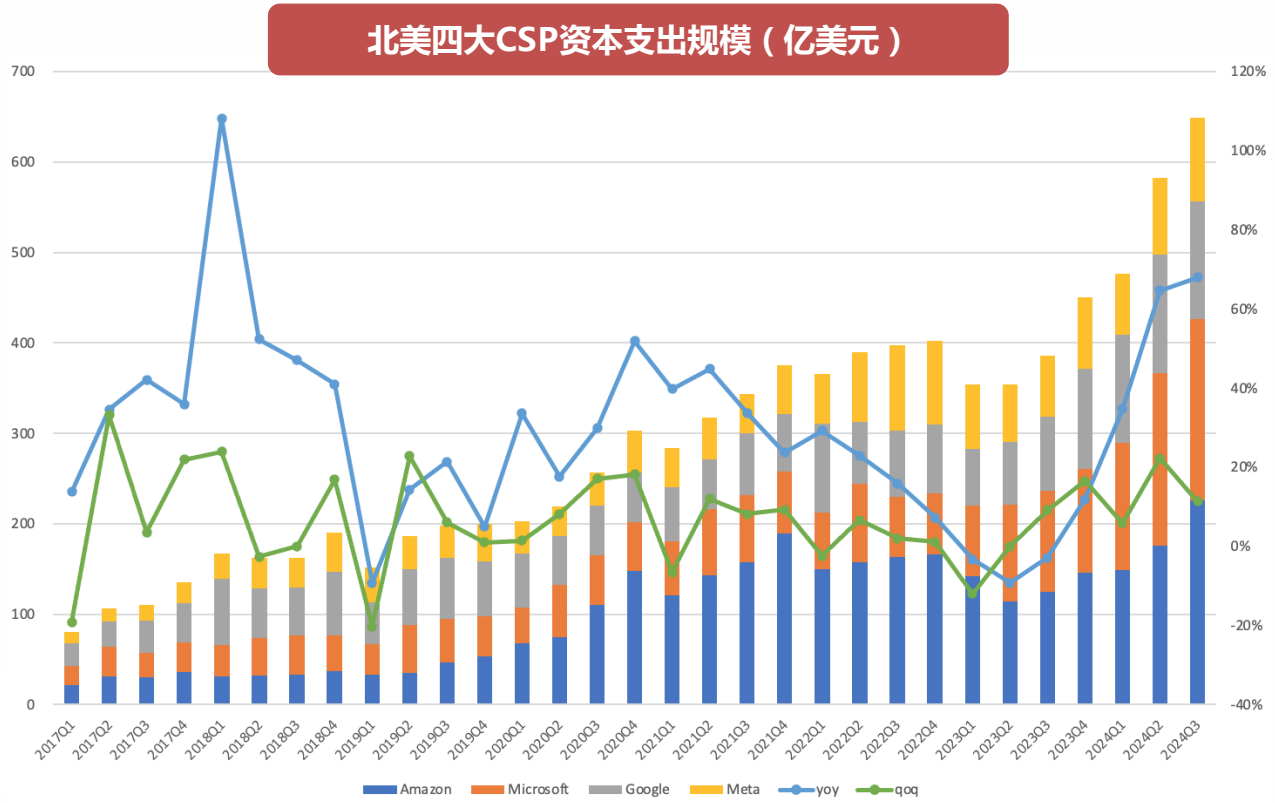
- **相关标的**：1）博通：全球AI ASIC龙头，目前已向多家头部CSP客户批量供应ASIC产品，其在计算，存储，网络IO，封装等领域广泛的IP储备可为其XPU产品线赋能。2）Marvell：全球一线ASIC厂商，其定制计算产品包括AI加速芯片，针对安全、NIC/DPU、ARM计算、存储、视频和CXL功能的ASIC等，客户包括北美头部云厂商。

- **风险提示**：AI产业发展不及预期的风险；大型科技企业资本支出不及预期的风险；GPU竞争的风险。

目 录

1.1 大型CSP加速资本支出

- 大型CSP在资本支出方面投入巨大，支出的同比增速在加快。北美四大CSP的Capex规模今年来增幅显著提升，2024年前三季度整体规模达到1708亿美元，同比增长56%，且yoy逐季加快（Q1-Q3 yoy分别为34.7%、64.6%、68%）。其中，微软530亿美元，yoy +78.5%；亚马逊551.7亿美元，yoy +44.6%；谷歌382.6亿美元，yoy +79%；Meta 243.9亿美元，yoy +20.7%。
- 资本支出大幅提升的背后，是各家巨头在AI赛道上的竞赛、AI算力的稀缺、AI云赋能和AI生态的拓展等多方面驱动。

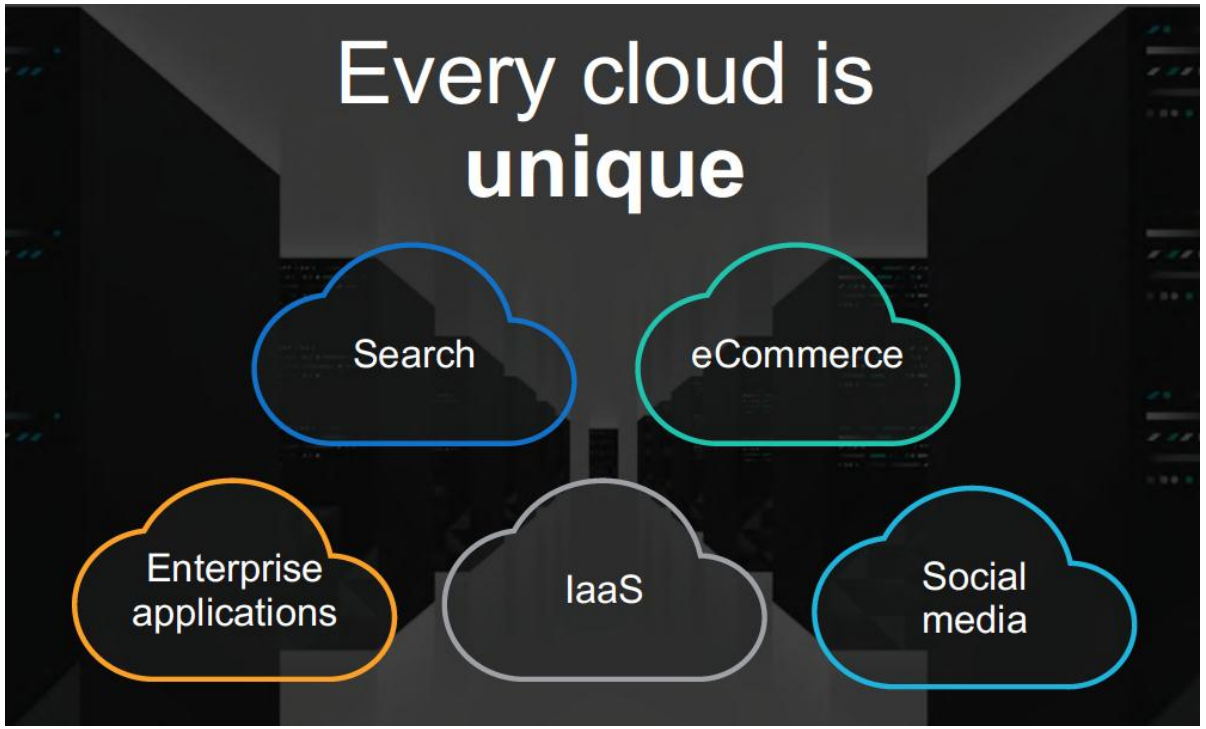


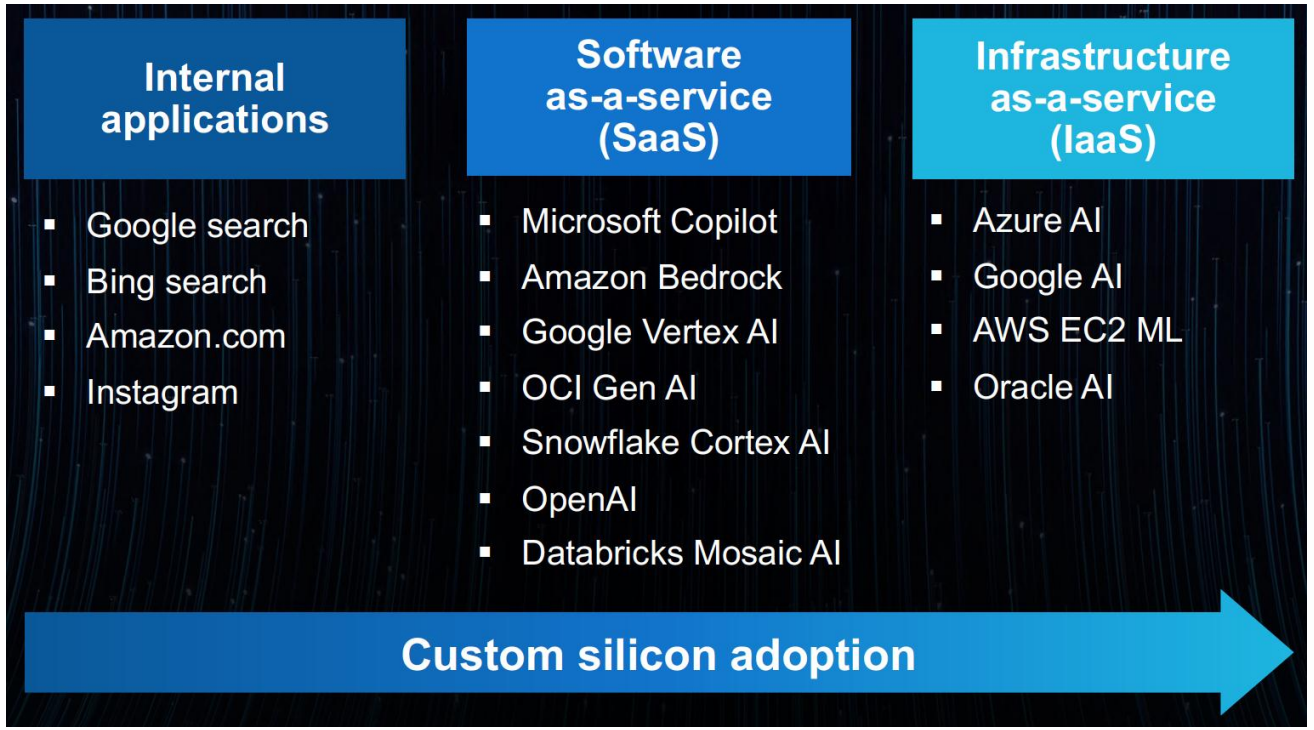
数据来源：各公司公告，西南证券整理

1.2 ASIC可适应不同的业务场景和商业模式的需求

- 大型CSP的业务模型、应用场景等很多通过自身的云来承载，每个云承载了独特的应用和商业模型，包括内部应用（比如搜索引擎、社交媒体等）、SaaS服务（比如AI聊天机器人、Copilot等）、IaaS服务等。ASIC 可以适应不同的业务场景和商业模式的需求。
- ASIC可以满足客户的需求：1）内部工作负载的架构优化；2）更低的功耗，更低的成本；3）为AI工作负载定制的内存和I/O架构。

ASIC需要满足不同业务/应用的加速计算需求



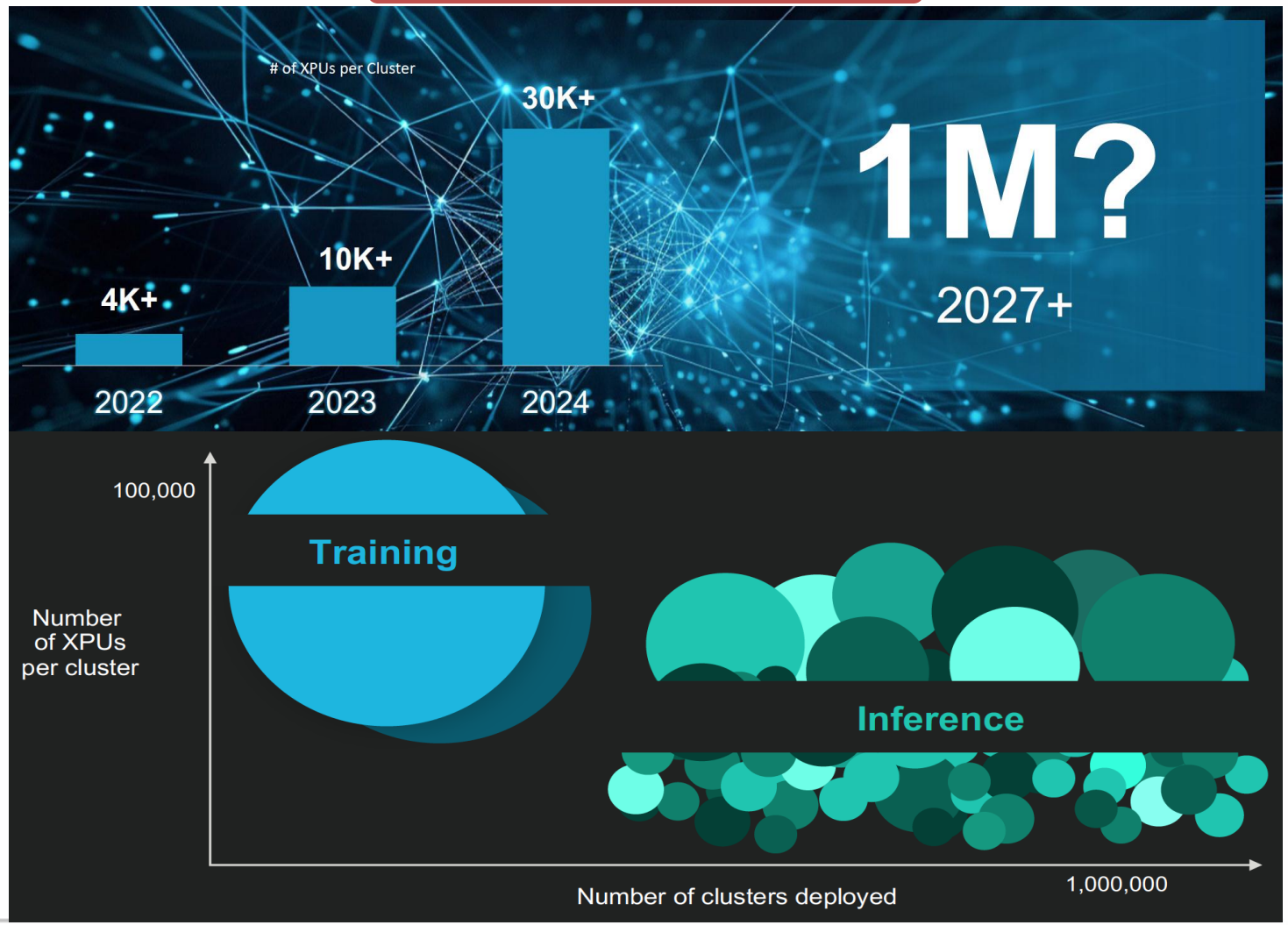


数据来源：Marvell，西南证券整理

1.3 训练和推理集群对加速计算芯片的需求

- 目前在训练阶段，训练集群对加速计算芯片的需求已提升到万卡级别。随着AI模型对训练需求的提升，未来10万卡级别指日可待。
- 而在推理阶段，由于计算量与业务和应用密切相关，单个推理集群对加速计算芯片的需求低于训练集群，但推理集群的部署数量要远多于训练集群，推理集群的数量预计会达到百万级别。
- AI算力集群特别是推理集群对加速计算芯片的庞大需求，是ASIC快速成长的核心驱动力。

训练和推理对AI算力集群的需求差异

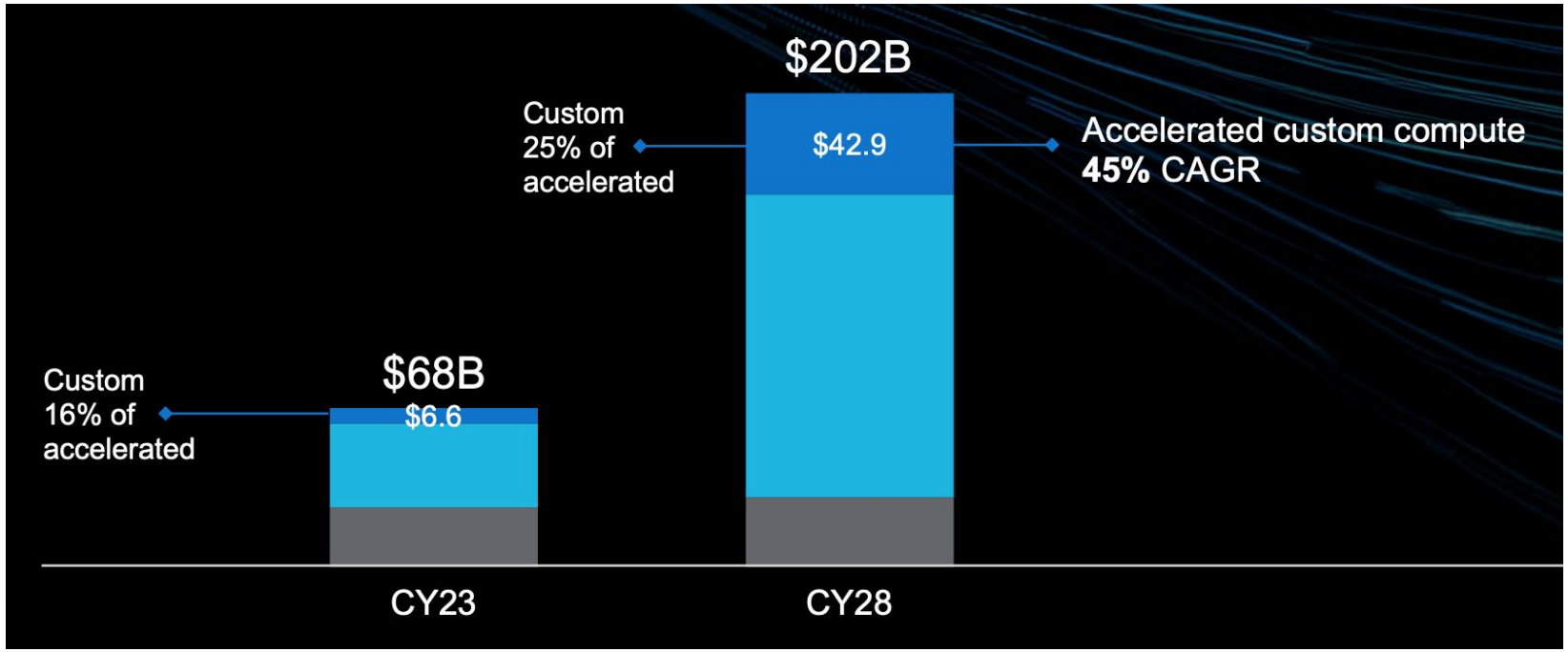


数据来源：Broadcom，Marvell，西南证券整理

1.4 ASIC市场规模预测

- 据Marvell预测，2023年 ASIC 占数据中心加速计算芯片的16%，规模约为66亿美元；随着 AI 计算需求的增长，ASIC 占比有望提升至25%，预计2028年数据中心 ASIC 市场规模将提升至429亿美元，CAGR为45.4%。

数据中心定制加速计算市场规模



数据来源：650 Group, CignalAI, Dell’Oro, LightCounting, Marvell, 西南证券整理

目 录

2.1 ASIC硬件性能：针对特定算法和应用优化设计，具有较高能效比

- **ASIC针对特定算法和应用进行优化设计，在特定任务上的计算能力强大**，例如在某些AI深度学习算法中实现高效的矩阵运算和数据处理。GPU具有强大的并行计算能力，拥有众多计算核心，可同时处理多个任务，在通用计算和图形处理方面表现出色，适用于大规模的数据并行计算，如科学计算、图形渲染、视频处理等；但GPU在特定任务上的计算效率可能不如ASIC。
- **ASIC通常具有较高的能效比，因其硬件结构是为特定任务定制的，能最大限度减少不必要的功耗。**GPU由于其通用的设计架构，在执行特定任务时可能存在一些功耗浪费；但随着技术的进步，新一代GPU也在不断提高能效比。
- **ASIC在处理特定任务时，能实现高吞吐量，数据处理速度快**，可快速完成大量的数据处理工作。GPU具有较高的带宽和并行处理能力，在图形处理和通用计算中能实现较高吞吐量，但在处理一些复杂、非图形相关的特定任务时，其吞吐量可能会受到一定限制。
- **ASIC在绝对算力和片间互联方面普遍低于AI GPU，但ASIC的服务器间互联由于采用以太网为主，具有通用性强、生态开放、低成本等优势。**

市面主流GPU与ASIC规格对比

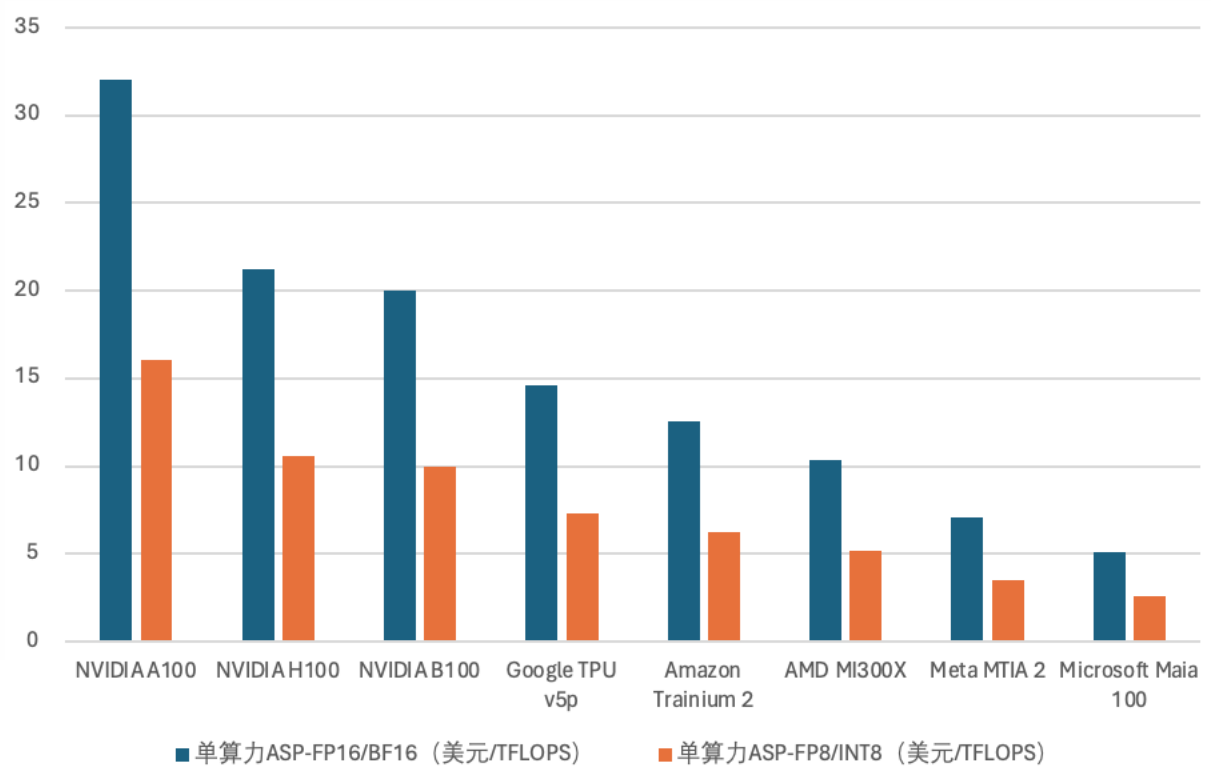
公司		英伟达		AMD		Google		亚马逊		微软	Meta
型号		H100	B100	MI300X	MI325X	TPU v5p	TPU v6 Trillium	Trainium 2	Inferentia 2	Maia 100	MTIA 2
算力（TFLOPS）	FP16/BF16	990	1750	1307	1307	459	926	430	190	800	354
	FP8/INT8	1979	3500	2615	2615	918	1852	860	380	1600	708
内存	类型	HBM3	HBM3e	HBM3	HBM3e	HBM2e	HBM3	HBM3	HBM2e	HBM2e	LPDDR5
	容量（GB）	96	200	192	256	96	96	96	48	64	128
	带宽（TB/s）	3.35	8	5.3	6	2.765	1.6	4	-	1.8	0.2
网络性能（GB/s）		900	1800	896	896	600	800	-	100	600	32
最大TDP（W）		700	700	750	1000	-	-	-	-	700	90
工艺制程（nm）		4	4	5	4	5	4	4	7	5	5

数据来源：各公司官网，西南证券整理

2.2 ASIC的单位算力成本更低，满足一定的降本需求

- **ASIC的单位算力成本更低，满足一定的降本需求。**ASIC因其硬件结构是为特定任务定制的，减少了很多针对通用加速计算的不必要的硬件设计，其单位算力成本相比GPU或更低。谷歌TPU v5、亚马逊Trainium 2的单位算力成本分别为英伟达H100的70%、60%

市面主流GPU与ASIC算力成本对比



数据来源：各公司官网，西南证券整理

2.3 ASIC与GPU软件生态对比

- ASIC在软件生态上的优势：**云厂商普遍具备较强的研发能力，为 ASIC 研发了配套的全栈软件生态，开发了一系列编译器、底层中间件等，提升 ASIC 在特定场景下的计算效率**。部分第三方芯片厂商推出了开源平台，未来 ASIC 的软件生态将会愈发成熟和开放。
- ASIC在软件生态上的劣势：软件生态相对较为单一，主要针对特定的应用场景和算法进行优化。与 GPU 相比，ASIC 的编程难度较大，需要专业的知识和技能，开发工具和软件库相对较少。这使得开发者在使用 ASIC 时需要花费更多时间和精力进行开发调试。
- GPU软件生态的优势：**软件生态丰富成熟，拥有广泛的开发工具、编程语言和软件库支持**，如英伟达的 CUDA 和 AMD 的 ROCm 等。开发者可使用熟悉的编程语言如 C、C++、Python 等进行开发，且有大量的开源项目和社区支持，方便开发者学习和交流。这使得 GPU 在各种应用场景中都能快速地进行开发和部署。
- GPU软件生态的劣势：软件生态在特定任务上的优化程度可能不如 ASIC。在一些对性能和功耗要求极高的特定场景中，需要进行大量的优化工作才能发挥出 GPU 的最佳性能。

数据来源：各公司官网，西南证券整理

2.4 ASIC以推理场景为主，并开始切入到部分训练环节

- ASIC在执行特定 AI 算法时的高性能和高能效的优势，对于大规模数据中心等对能耗敏感的场景非常重要。由于 ASIC 不需要集成通用的功能模块，从而减少不必要的硬件资源浪费，如果AI应用场景明确且需求量大，ASIC在大规模生产后其单位成本可显著降低。但ASIC也有开发周期长且灵活性差的劣势，由于ASIC的设计和制造是针对特定算法和应用场景进行的，一旦设计完成其功能就固化下来，难以对芯片的功能和性能进行修改和升级，如果 AI 算法发生较大变化，ASIC 可能无法快速适应这种变化。此外，ASIC的生态系统还不够完善，开发者在使用 ASIC 时可能需要花费更多时间和精力去搭建开发环境、编写底层代码等，开发难度较大。
- ASIC更适用于推理：在推理阶段，AI模型已训练完成，需要对输入的数据进行快速的预测和分类。此时对芯片的计算精度要求相对较低，但对计算速度、能效和成本等要求较高。ASIC正好满足这些需求，其高度定制化的设计能针对推理任务进行优化，以较低的功耗实现快速的推理计算。且在大规模部署的场景下，ASIC的成本优势更加明显，可以降低企业的运营成本。
- GPU更适用于训练：AI训练过程需要处理大量的数据和复杂的计算，对芯片的计算能力、内存带宽和并行处理能力要求非常高。GPU拥有众多的计算核心和高带宽内存，可以同时处理大量的数据样本和复杂的计算任务，能够加速 AI 模型的训练过程。且在训练过程中，需要不断地调整模型的参数和结构，GPU的灵活性使其更适合这种频繁的调试和迭代。

数据来源：西南证券整理

目 录

3.1 谷歌TPU：谷歌专为AI定制设计的ASIC

- **TPU（Tensor Processing Units，张量处理单元）是谷歌专为AI定制设计的ASIC，其针对大模型的训练和推理进行了优化。**TPU适合各种使用场景，例如聊天机器人、代码生成、媒体内容生成、合成语音、视觉服务、推荐引擎、个性化模型等。
- 截至2024年，谷歌TPU已迭代6代产品。**TPU v5p单个Pod可达8960颗芯片的集群规模，借助Multislice训练技术，TPU v5p可实现5万卡线性加速。最新一代TPUv6 Trillium预计2024H2推出，TPU v6 FP16/BF16精度非稀疏算力可达926 TFLOPS，约为H100、B100的93%、53%。相比TPU v5e，TPU v6能效高出67%，峰值性能高出3.7倍。**

谷歌TPU历代产品性能

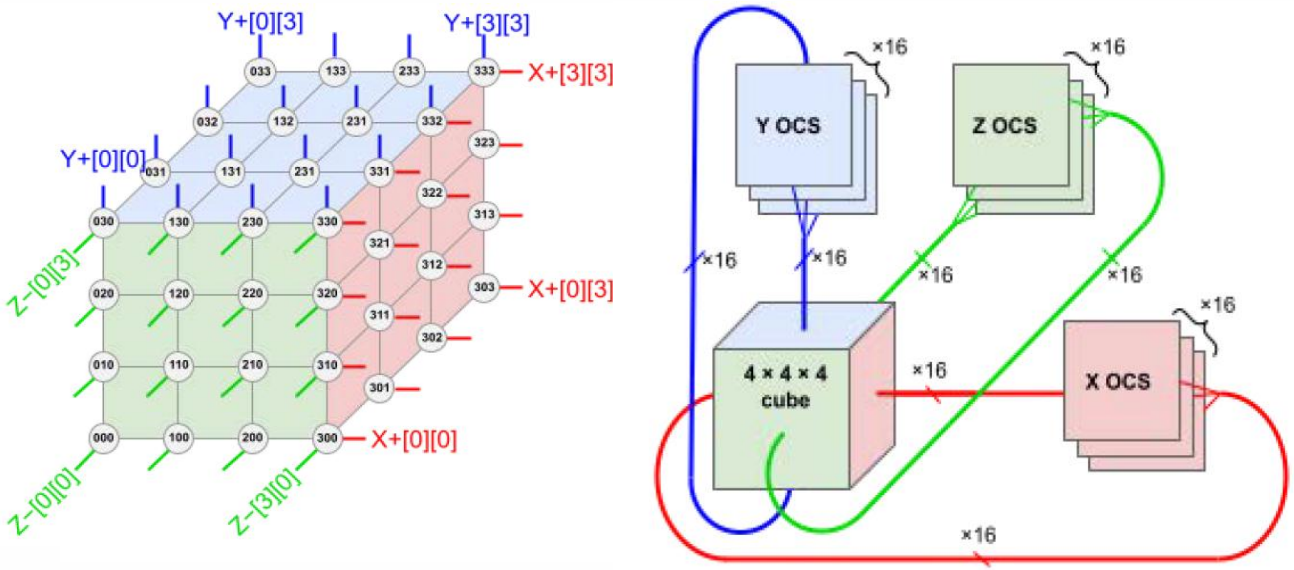
产品	v1	v2	v3	v4	v5e	v5p	v6
单Pod芯片数量	-	256	1024	4096	256	8960	4096
FP16/BF16算力 (TFLOPS)	-	46	123	275	197	459	926
INT8算力 (TFLOPS)	92	92	246	-	394	918	1852
HBM容量 (GB)	8	16	32	32	16	96	96
HBM带宽 (GB/s)	34	700	900	1200	819	2765	1640
工艺制程 (nm)	28	16	16	7	5	5	5

数据来源：谷歌，西南证券整理

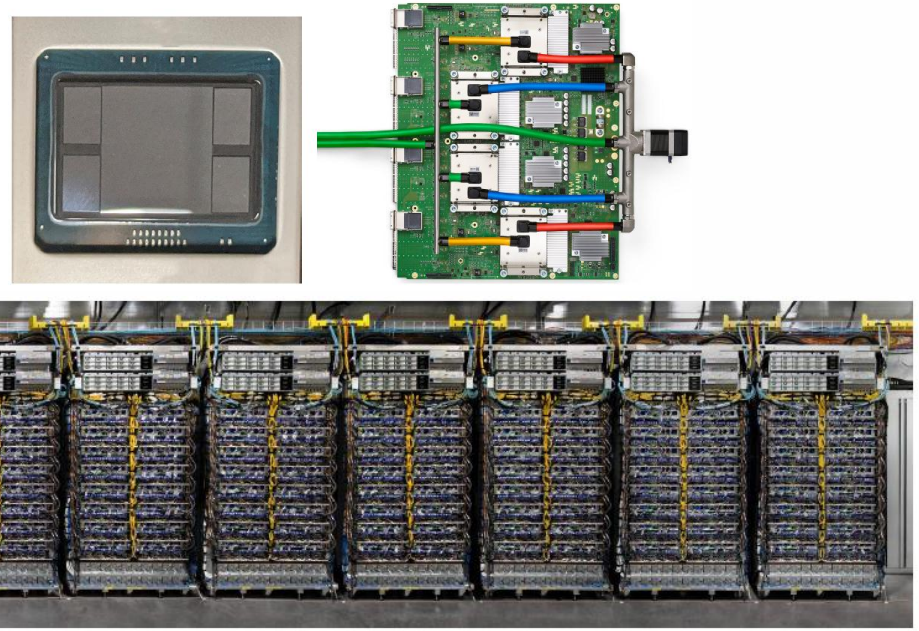
3.1.1 谷歌TPU算力集群能力

- TPU v4和TPU v5p算力集群采用3D torus(3D环面)架构和OCS，提供高速的网络连接，增强拓展性与互联效率。在TPUv4的架构中，每64颗TPU v4芯片组成4x4x4的立方体，每个CPU配备4颗TPU v4，64颗TPU v4和16颗CPU放入一个机架，形成一个模块。
- 一个模块有6个面的光路链接，每个面有16个链接，单模块共有96个光路连接到OCS。为了提供3D环面的环绕链接，对面的链接必须连接到同一个OCS。每个模块连接48个OCS（6×16÷2），最终实现所需的4096个TPU v4芯片互联。
- TPU v4算力集群的物理架构：一个PCB包含4个TPU v4，通过ICI链路连接到其他托盘（tray），16个托盘共同放入一个机架，形成4x4x4的3D模块结构。64个机柜共同组成4096颗芯片规模的算力集群。

TPU立方体与3个OCS的连接



TPU v4封装和算力集群

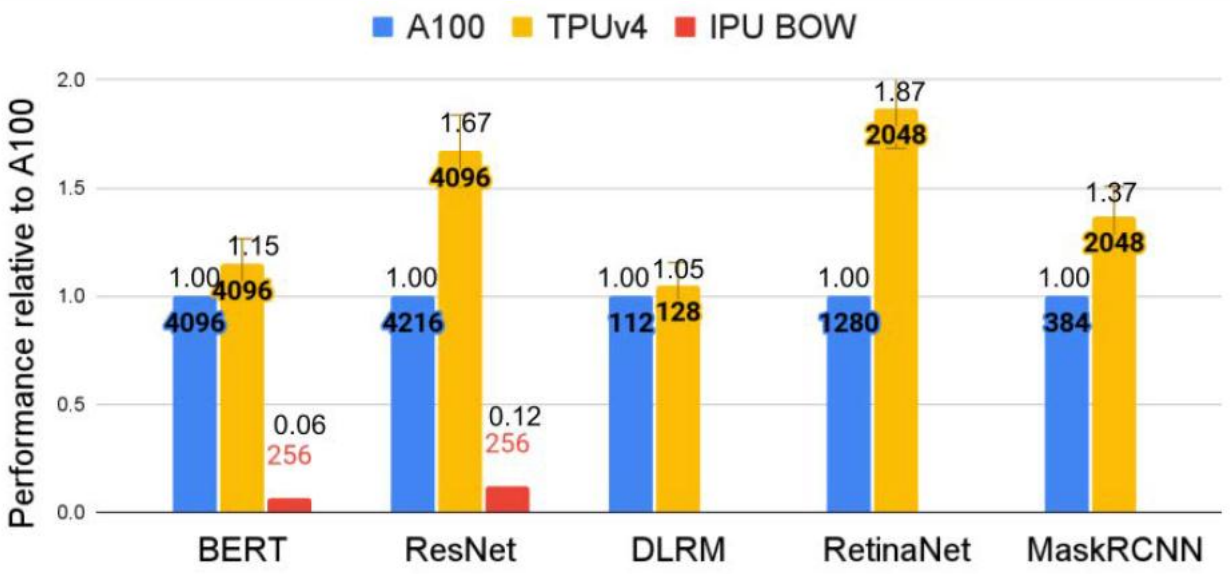


数据来源：谷歌，《TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings》，西南证券整理

3.1.2 谷歌TPU基准测试性能对比

- TPU v4与英伟达A100在MLPerf基准测试中的性能对比：TPU v4在BERT上比A100快1.15倍，比IPU快约4.3倍；在ResNet上，TPU v4分别比A100和IPU快1.67倍和约4.5倍；运行MLPerf基准测试时，A100的平均功耗比TPU v4高1.3~1.9倍。虽然TPU v4单芯片算力为A100的88%，但在性能和功耗表现上要优于A100。

TPU与A100在MLPerf训练中的性能对比



TPU与A100在MLPerf训练中的功耗对比

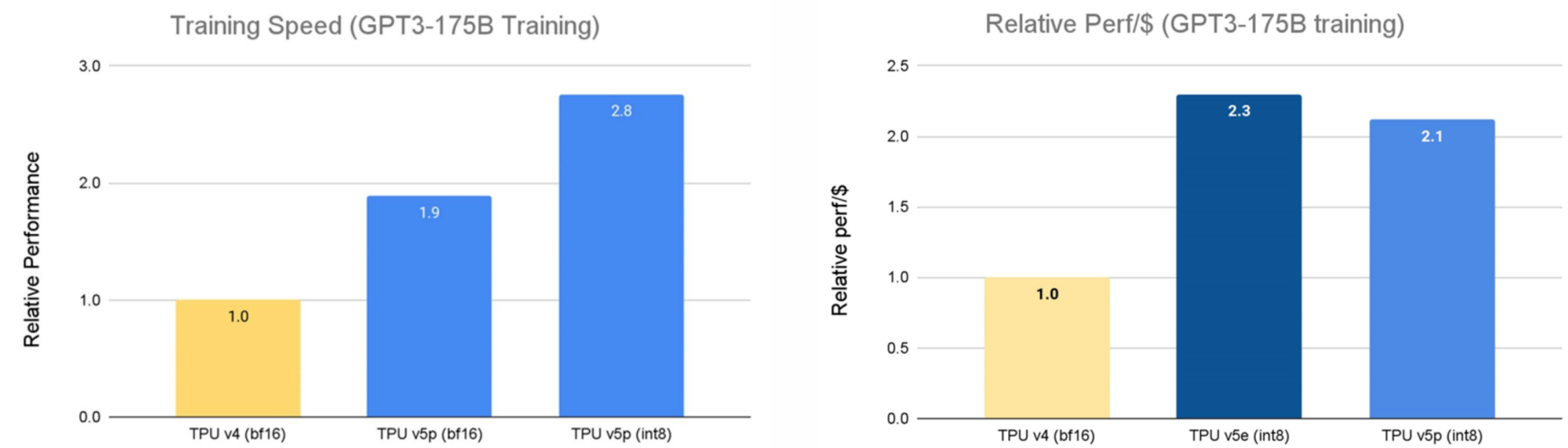
MLPerf Benchmark	英伟达A100	谷歌TPU v4	比率
BERT	380 W	197 W	1.93
ResNet	273 W	206 W	1.13

数据来源：谷歌，《TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings》，西南证券整理

3.1.3 谷歌TPU迭代推动大模型训练效率显著提升

- TPU的算力成本随着产品更新迭代也在持续优化。**TPU v5e的相对性价比（TFLOPs/$）是TPU v4的2.3倍，参考谷歌披露的TPU v4公开标价3.22美元/芯片/小时，TPU v5e的标价为1.2美元/芯片/小时，TPU v5e以更低的成本实现了更高的算力。TPU v5p训练LLM的速度比TPU v4快2.8倍，利用第二代SparseCores，TPU v5p训练嵌入密集模型的速度比TPU v4快1.9倍。**

谷歌TPU迭代推动大模型训练效率的显著提升



数据来源：谷歌，西南证券整理

3.2.1 亚马逊自研AI芯片Trainium

- **AWS Trainium是AWS专门为超过1000亿个参数模型的深度学习训练打造的机器学习芯片。**自2020年以来，亚马逊发布了两代Trainium芯片。Trainium 1加速器提供190 TFLOPS的FP16/BF16算力，配有32GB的HBM，内存带宽820GB/s；而新一代Trainium 2达到了430 TFLOPS的FP16/BF16算力，其HBM容量达到96GB，内存带宽为4TB/s。**与第一代相比，AWS Trainium 2的性能提高了4倍，能效提高了1倍。**
- **每个Amazon Elastic Compute Cloud (Amazon EC2) Trn1实例部署多达16个Trainium加速器。AWS表示未来扩展到多达10万个芯片的EC2 UltraCluster集群中**，从而高效训练大模型。基于Trainium的Amazon EC2 Trn1实例与同类Amazon EC2实例相比，可节省高达50%的训练成本。Trainium已针对训练自然语言处理、计算机视觉和推荐器模型进行了优化，这些模型用于文本摘要、代码生成、问题解答、图像和视频生成、推荐和欺诈检测等各种应用程序。

AWS Trainium 2

AWS Trainium产品线

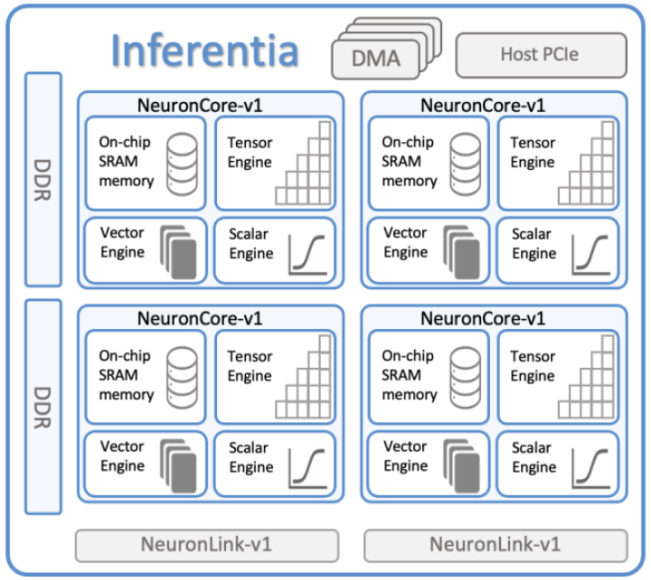
产品	Trainium 1	Trainium 2
Core类型	NeuronCore-v1	NeuronCore-v2
内核数量	32	64
NeuronLink带宽 (GB/s)	768	2043
FP16/BF16算力 (TFLOPS)	190	431
INT8算力 (TFLOPS)	380	861
HBM容量 (GB)	32	96
HBM带宽（GB/s）	820	4000
晶体管数量（亿）	550	1150

数据来源：AWS，Semianalysis，西南证券整理

3.2.2 亚马逊自研AI芯片Inferentia

- **AWS Inferentia加速器由AWS设计，在Amazon EC2中以低成本为深度学习和生成式AI推理应用程序提供高性能。**第一代AWS Inferentia 1加速器为Amazon Elastic Compute Cloud（Amazon EC2）Inf1实例提供支持，与同类Amazon EC2实例相比，该实例的吞吐量可提高多达2.3倍，每次推理的成本可降低多达70%。
- 2023年亚马逊发布了Inferentia 2芯片和Inf2实例，与Inferentia相比，AWS Inferentia 2加速器的吞吐量提高了4倍，延迟低至上一代的1/10。Inferentia 1加速器搭载4个第一代NeuronCore，配有8 GB的DDR4内存，每个EC2 Inf1实例最多有16个Inferentia 1加速器。Inferentia 2加速器搭载了2个第二代NeuronCore，支持190 TFLOPS的FP16性能，配置32GB的HBM，与Inferentia 1相比，总内存增加了4倍，内存带宽增加了10倍；每个EC2 Inf2实例最多有12个Inferentia 2加速器。

AWS Inferentia 1架构



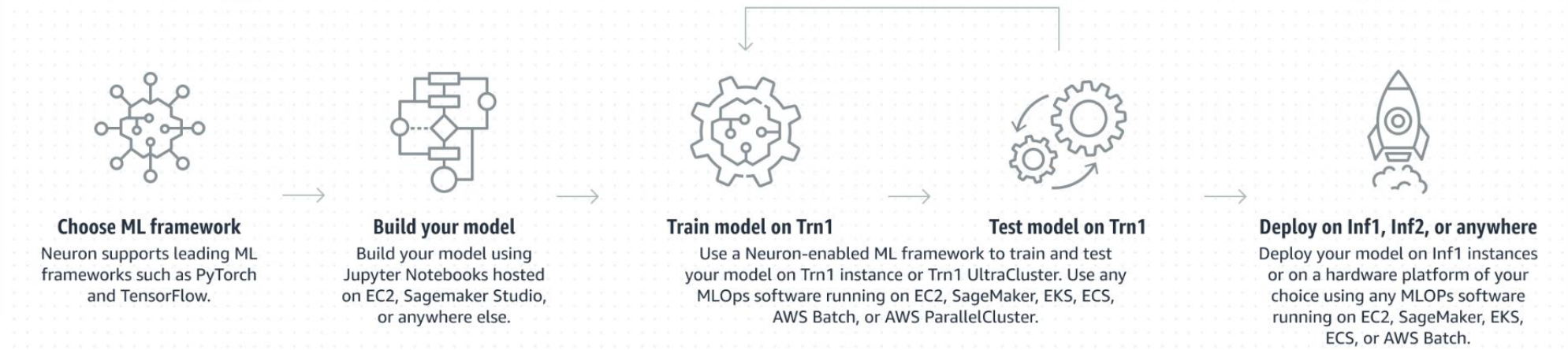
AWS EC2 Inf2实例性能

数据来源：AWS，西南证券整理

3.2.3 亚马逊AWS Neuron

- **AWS Neuron是一款用于优化AWS Trainium和AWS Inferentia加速器上的机器学习性能的SDK。**它支持在基于AWS Trainium的Amazon EC2 Trn1 实例上进行高性能训练。对于模型部署，它支持在基于AWS Inferentia的Amazon EC2 Inf1实例和基于AWS Inferentia2的Amazon EC2 Inf2实例上进行高性能和低延迟推理。AWS Neuron SDK与PyTorch和TensorFlow原生集成，确保客户可继续在这些热门框架中使用现有工作流程，并在Amazon EC2 Trn1、Inf1和Inf2实例上以最佳方式训练和部署ML/DL模型。开发者可将基于GPU的实例迁移到AWS Tranium中，客户只要修改少量代码即可实现海量数据训练，降低了训练成本。

AWS Neuron SDK工作流程



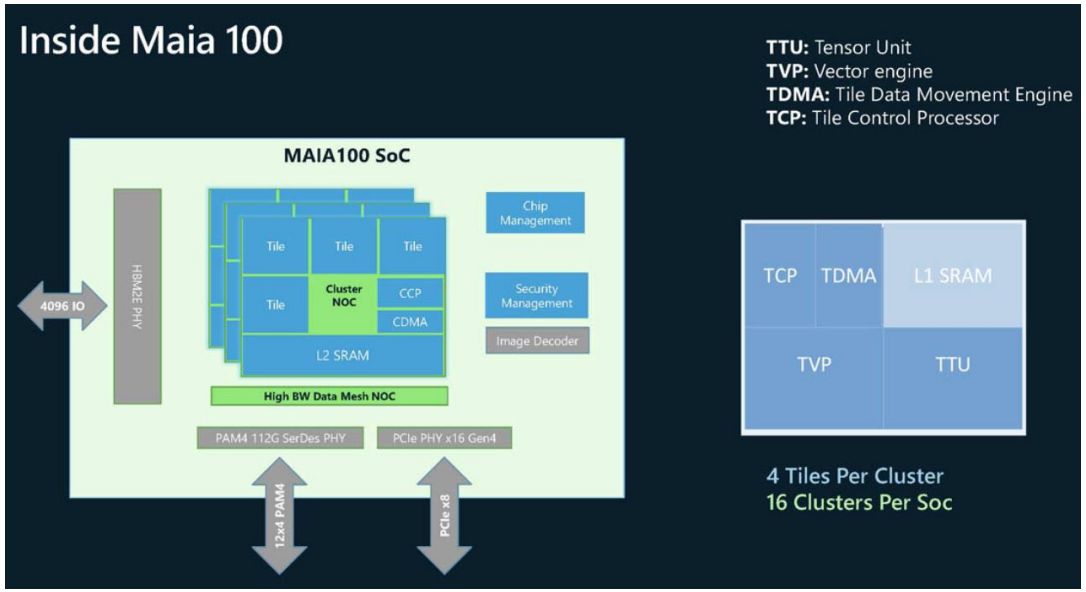
数据来源：AWS，西南证券整理

3.3 微软自研芯片Maia 100

- 微软将Maia 100打造成定制的AI加速器，用于在Azure上运行OpenAI的模型和Copilot等AI工作负载。**Maia 100采用台积电5nm制程和CoWoS-S封装技术，配备64GB（4×16GB）的HBM2E，内存带宽达1.8TB/s。**Maia 100配备一个500MB的L1/L2缓存，芯片具有12倍400GbE的网络带宽，设计最大功耗700W TDP。
- **Maia 100芯片在MXFP4数据格式下的性能达到 3200 TFLOPS，Int8下达到 1600 TFLOPS，BF16下达到 800TFLOPS，算力性能超过英伟达A100 28%，是英伟达H100的40%。**
- 微软Maia 100单SoC搭载16个集群，其中每个集群搭载4个图块Tile。Maia 100拥有图像解码器和机密计算能力，支持广泛的数据类型，包括FP32和BF16。

微软Maia 100介绍

微软Maia 100内部结构

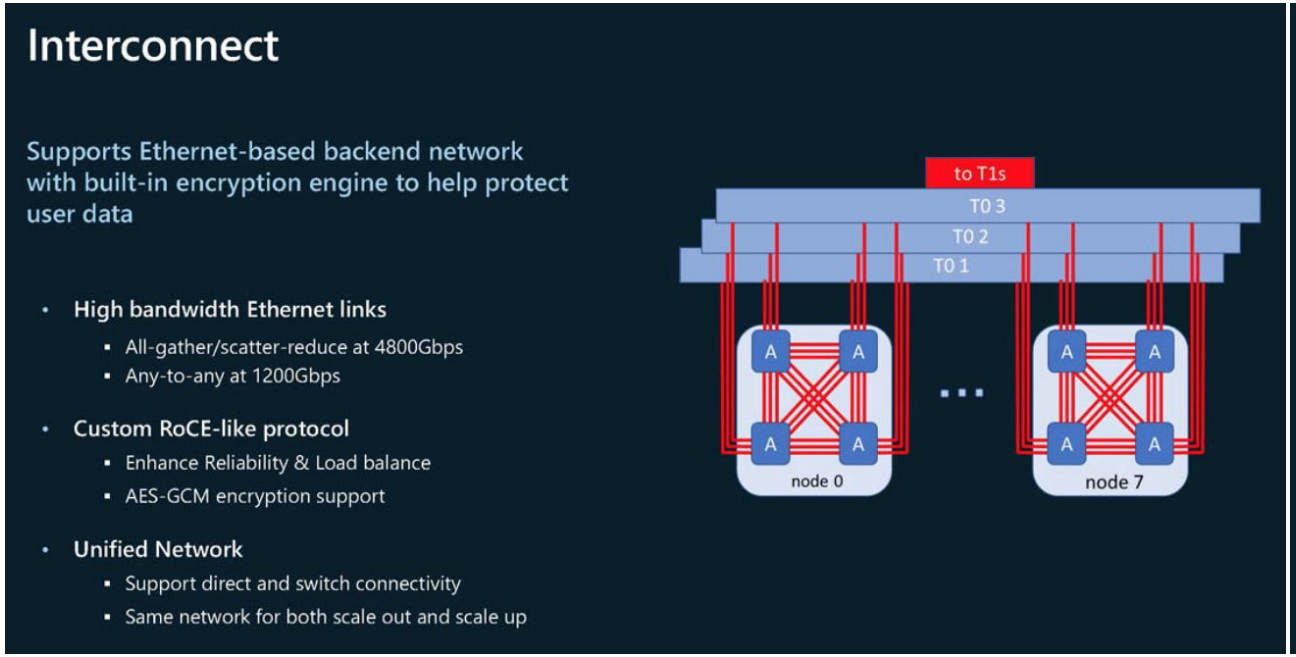


数据来源：微软，西南证券整理

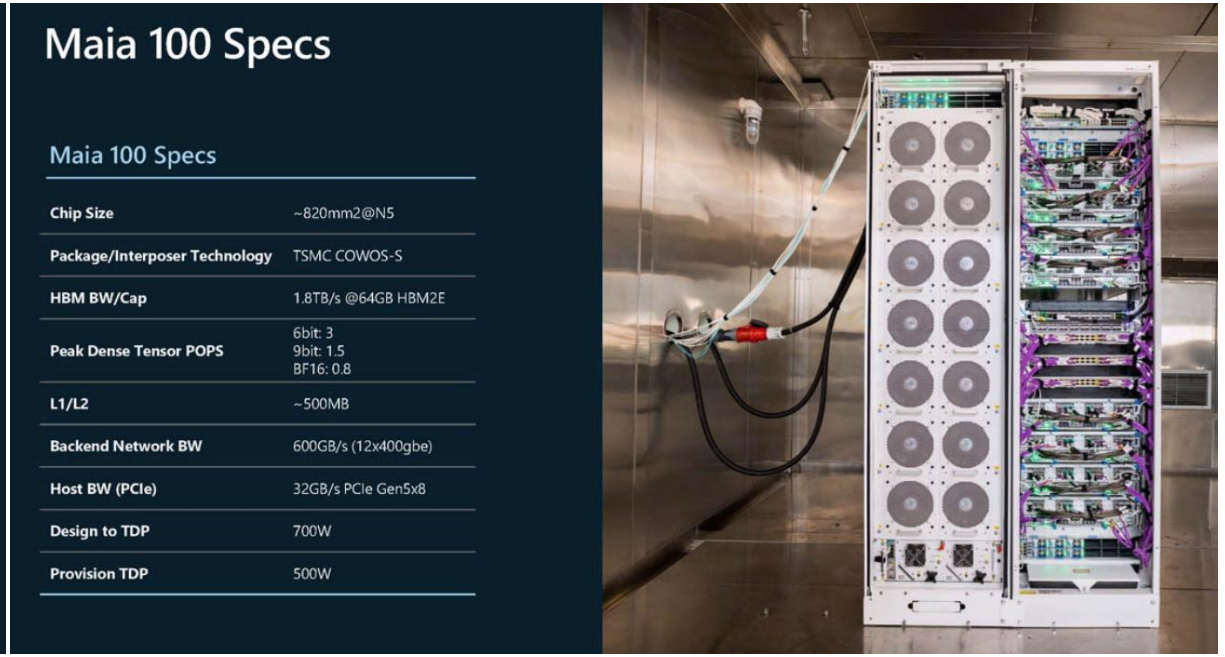
3.3 微软自研芯片Maia 100

- Maia 100基于自定义的RoCE类协议和以太网互连，内置AES-GCM加密引擎以保护用户数据，网络连接带宽达到600GB/s。Maia 100还由统一的后端网络支持，用于扩展和横向扩展工作负载，提供了支持直接和交换机连接的灵活性。
- 微软Maia 100芯片的Ares机架配备32颗Maia 100。**Ares一个机架中搭载了8台服务器，每台服务器中含有4个Maia 100，因此一个机架中总共有32颗Maia 100芯片**。Ares机架功率可达40kW，配置了Sidekick液体冷却系统，在机架两侧设置副设备，冷液从副设备流向Maia 100表面的冷板，副设备吸取液体中热量后再将冷液输出，以此构建散热循环。

微软Maia 100以太网络拓扑



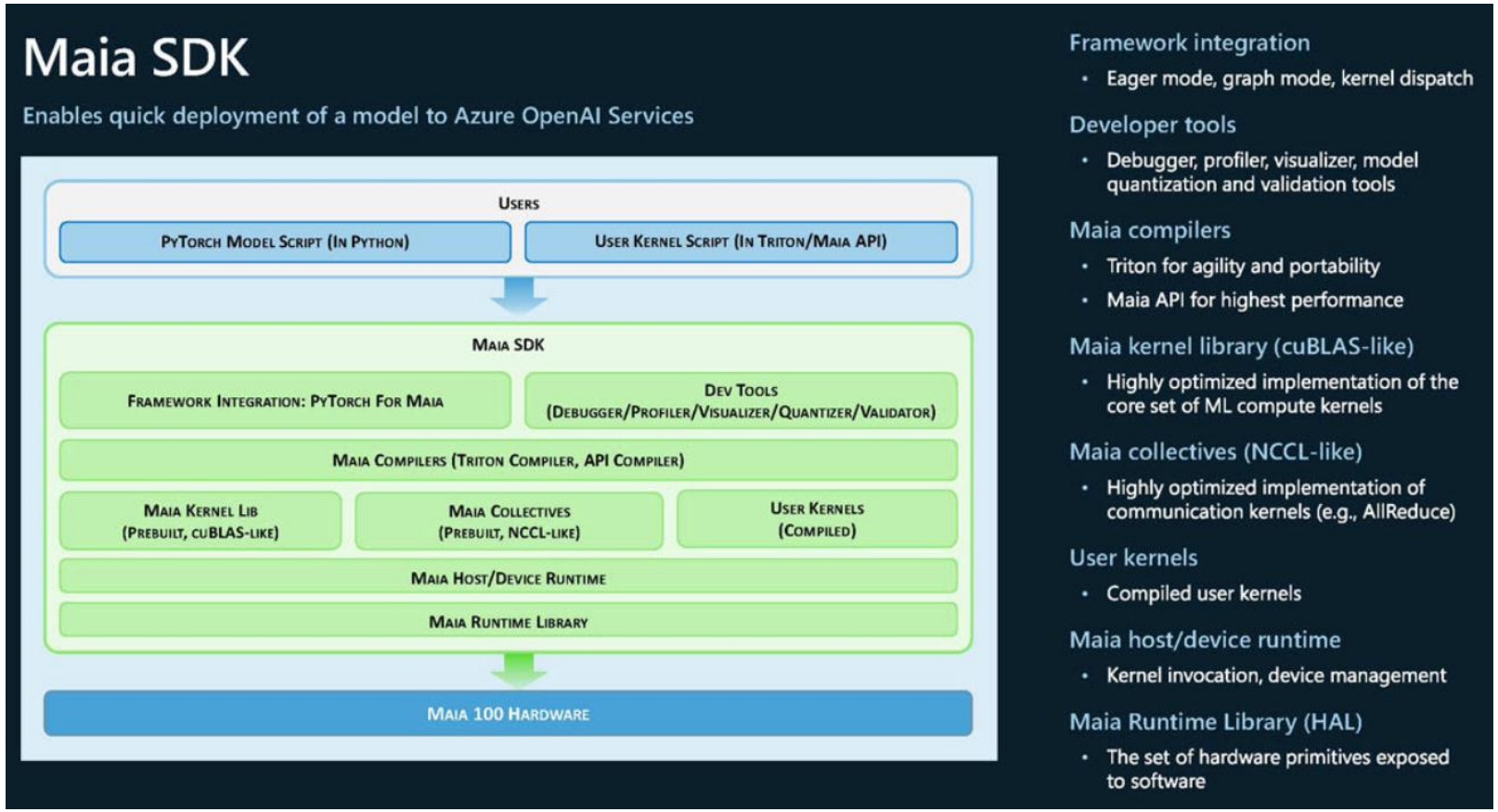
微软Maia 100规格和Ares机架



数据来源：微软，西南证券整理

3.3 微软自研芯片Maia 100

- **Maia SDK上实现快速部署和模型可移植性**。微软为Maia 100创建了软件，该软件与PyTorch和ONNX Runtime等流行的开源框架集成。该软件栈提供了丰富而全面的库、编译器和工具，使数据科学家和开发人员能在Maia 100上成功运行模型。微软集成了OpenAI的Triton；Triton是一种开源编程语言，通过抽象底层硬件简化了内核编写，这将赋予开发者完全的可移植性和灵活性，而不会牺牲效率和针对AI工作负载的能力。Maia的SDK允许用户将用PyTorch和Triton编写的模型快速移植到Maia。

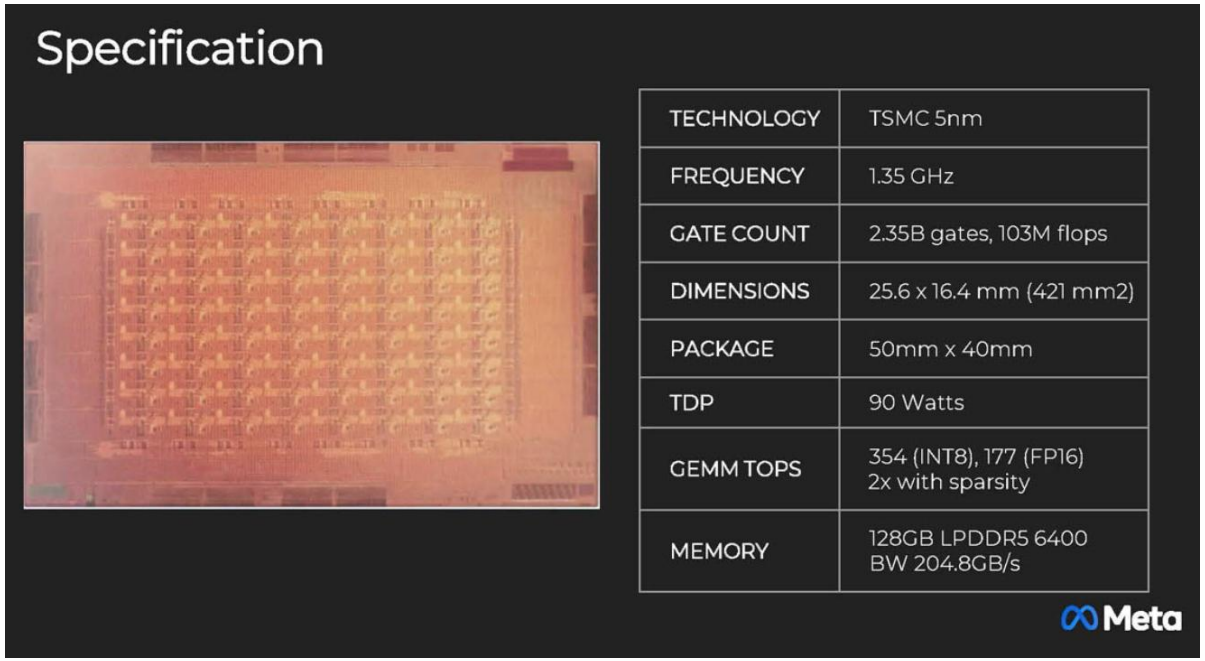


数据来源：微软，西南证券整理

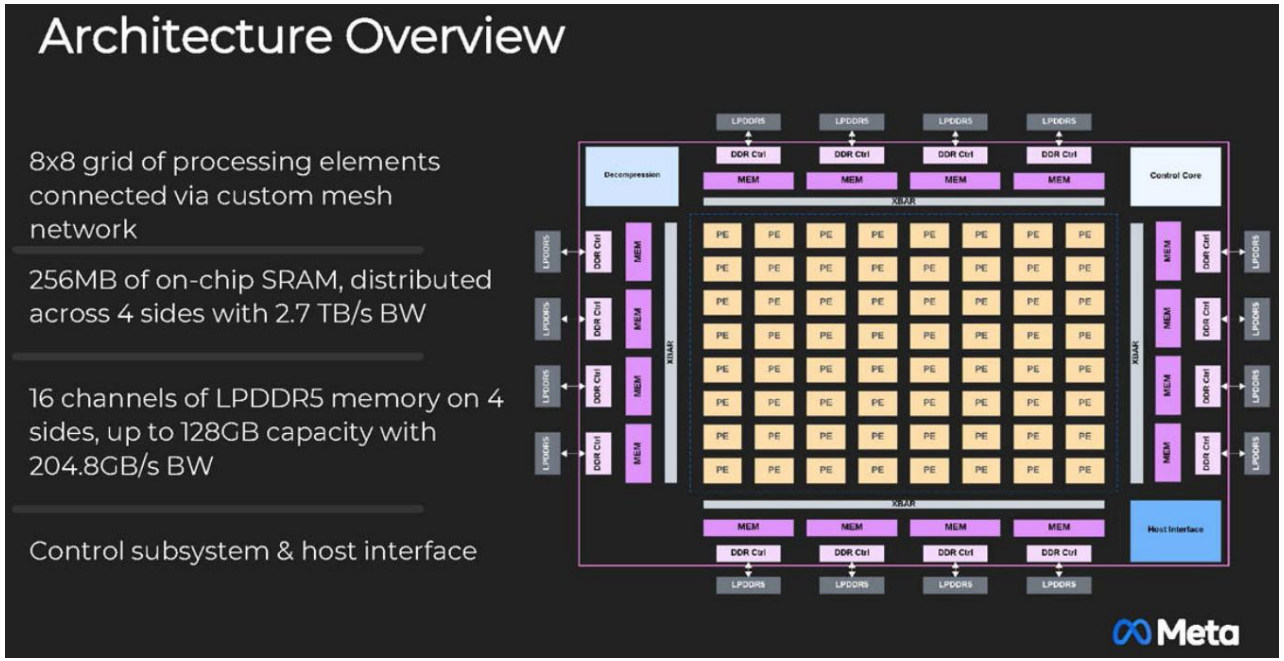
3.4 Meta自研芯片MTIA

- MTIA v2于2024年4月发布，**用于AI推理，旨在增强Meta的排名和广告推荐引擎**。MTIA v2采用台积电5nm制程，与上一代相比算力和内存带宽翻倍提升，INT8下的稠密算力354 TFLOPS接近上一代的3.5倍，稀疏算力708 TFLOPS达到上一代的近7倍。MTIA v2配备128GB的LPDDR5内存，内存带宽205GB/s，设计最大功耗90W TDP 。目前Meta已有16个数据中心使用了新款芯片。
- 芯片架构方面，MTIA v2内部包含加速器、片上和片外存储以及互联结构。AI加速器由8x8的处理单元网格（PE，processing element）组成，PE基于RISC-V内核，PE彼此互联，可作为一个整体运行任务，也可以独立处理任务。片上内存SRAM容量256MB，SRAM带宽为2.7TB/s，每个PE内存容量为384KB，PE带宽为1 TB/s。每个加速器使用PCIe Gen5 x8主机接口。

Meta MTIA 2规格



Meta MTIA 2芯片架构

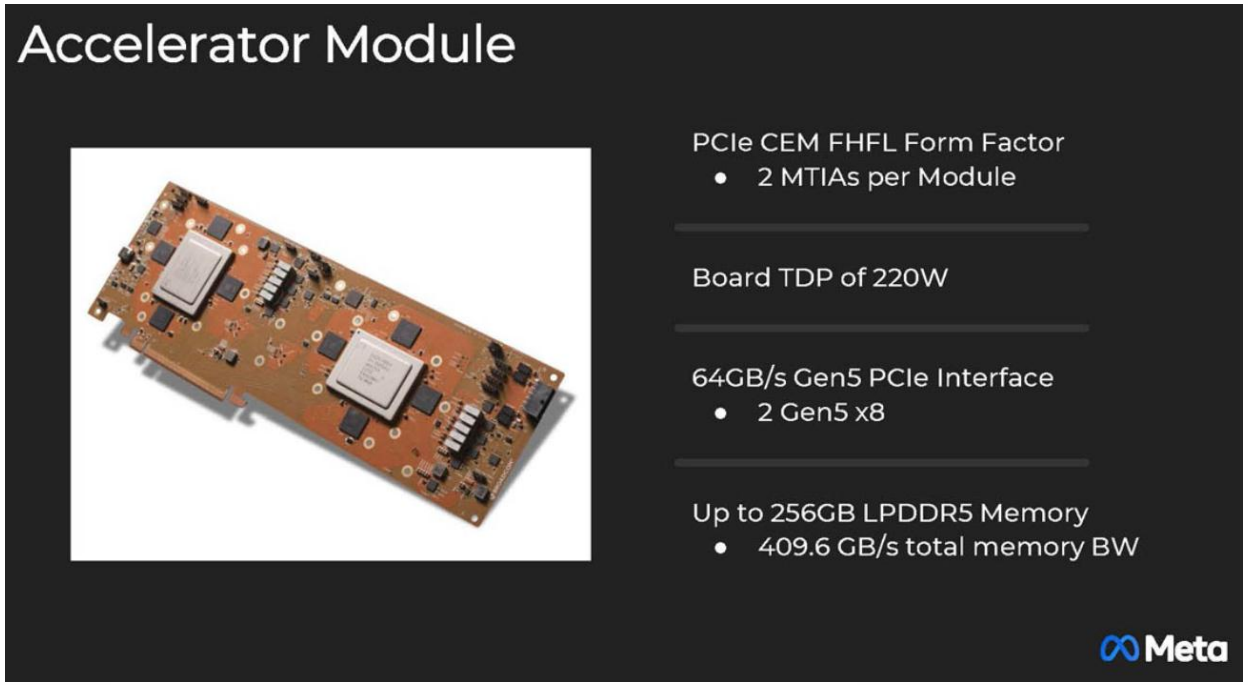


数据来源：Meta，西南证券整理

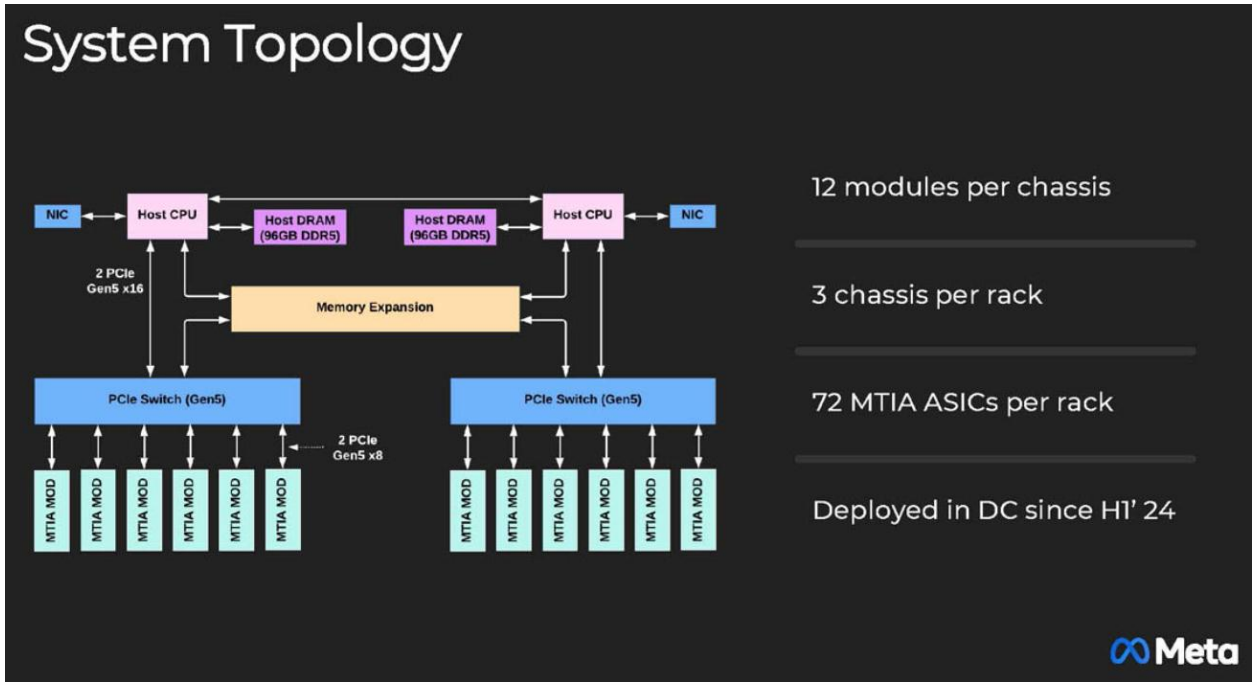
3.4 Meta自研芯片MTIA

- MTIA v2加速器模块：每张卡2个MTIA芯片，每个MTIA都可以使用PCIe Gen5 x8接口，单模块共x16接口（2 PCIe Gen5 x16）。
- MTIA机柜系统结构：一个机架系统包含(2×MTIA芯片)×(12×模组)×(3×机箱)，相当于每个机架系统搭载了72颗MTIA芯片。

Meta MTIA 2加速器模块结构和性能指标



Meta MTIA 2机架系统拓扑

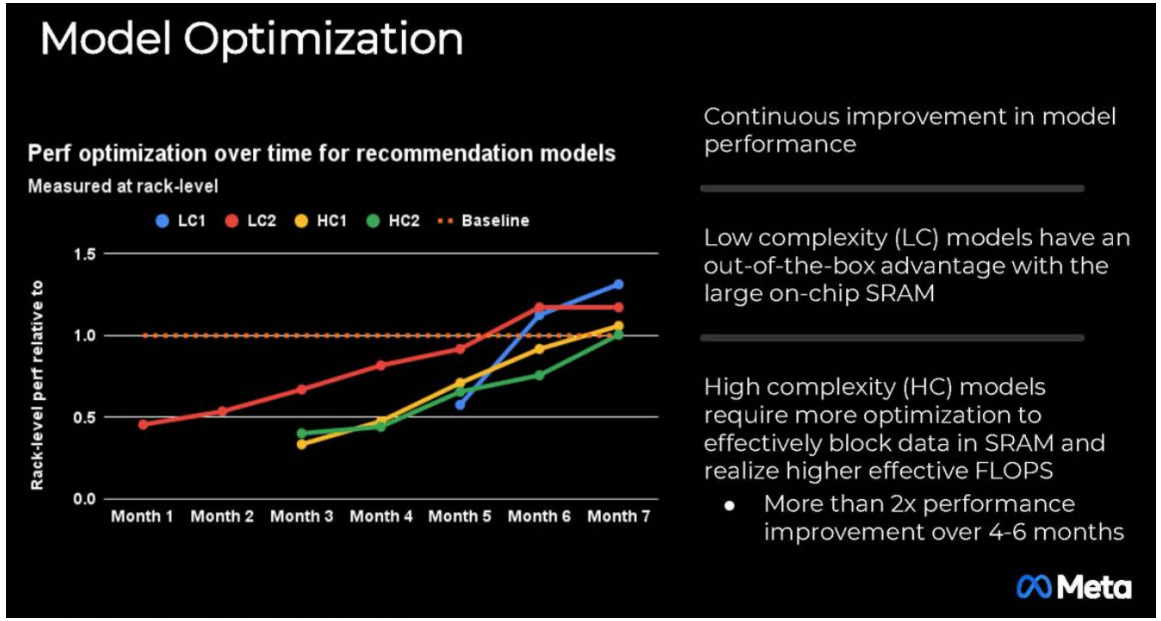


数据来源：Meta，西南证券整理

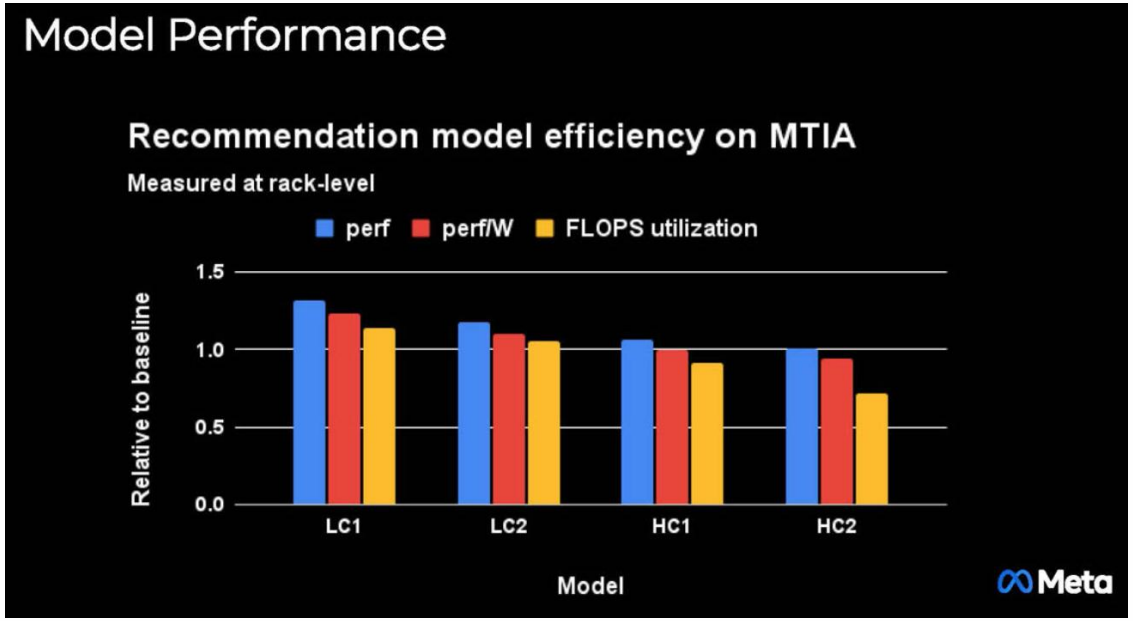
3.4 Meta自研芯片MTIA

- MTIA v2软件堆栈与PyTorch 2.0、TorchDynamo、TorchInductor完全集成，致力于提高开发者编程效率。MTIA v2的低级编译器从前端获取输出，生成高效且特定于设备的代码。下方是运行时堆栈，负责与驱动程序/固件接口，最后，运行时与驱动程序交互。Meta创建了Triton-MTIA编译器后端为芯片硬件生成高性能代码，Triton用于编写ML计算内核，极大提高了开发人员效率。
- 基于MTIA平台加速后的Meta推荐模型的效率得到提升，在大型片上SRAM的加持下，低复杂度（LC）模型具有开箱即用的优势，而高复杂度（HC）模型在4-6个月内性能提高了2倍以上。

Meta内部工作负载的模型性能基线



基于MTIA平台加速后的Meta推荐模型效率得到提升



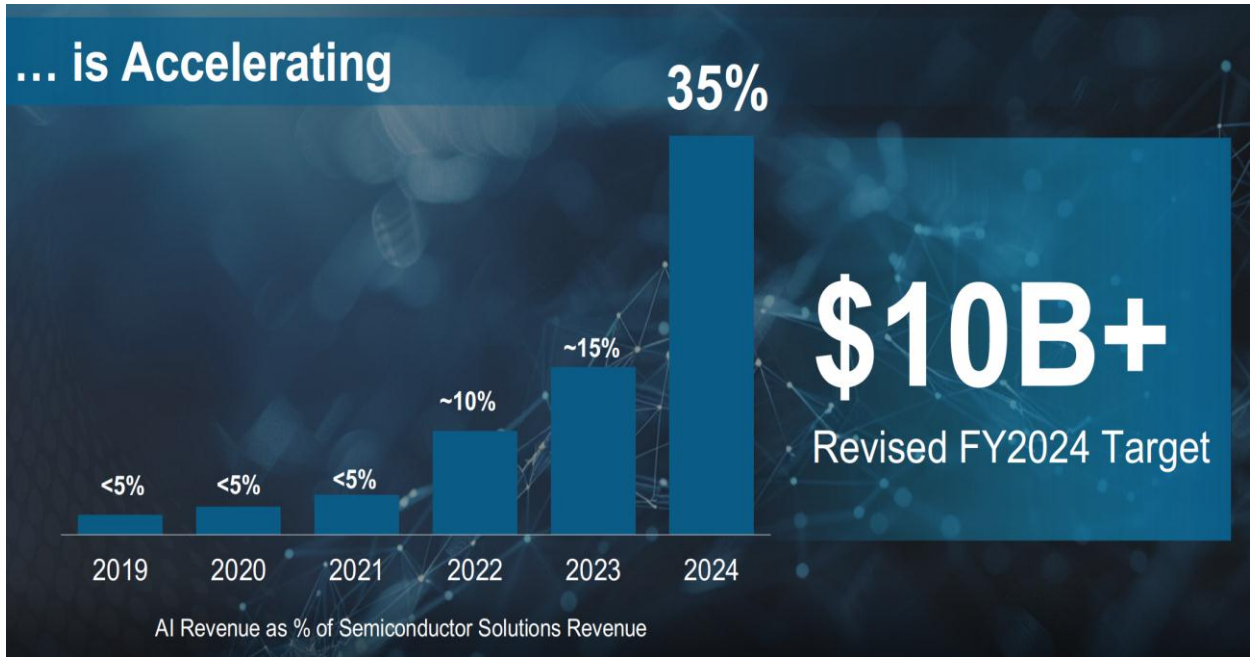
数据来源：Meta，西南证券整理

目 录

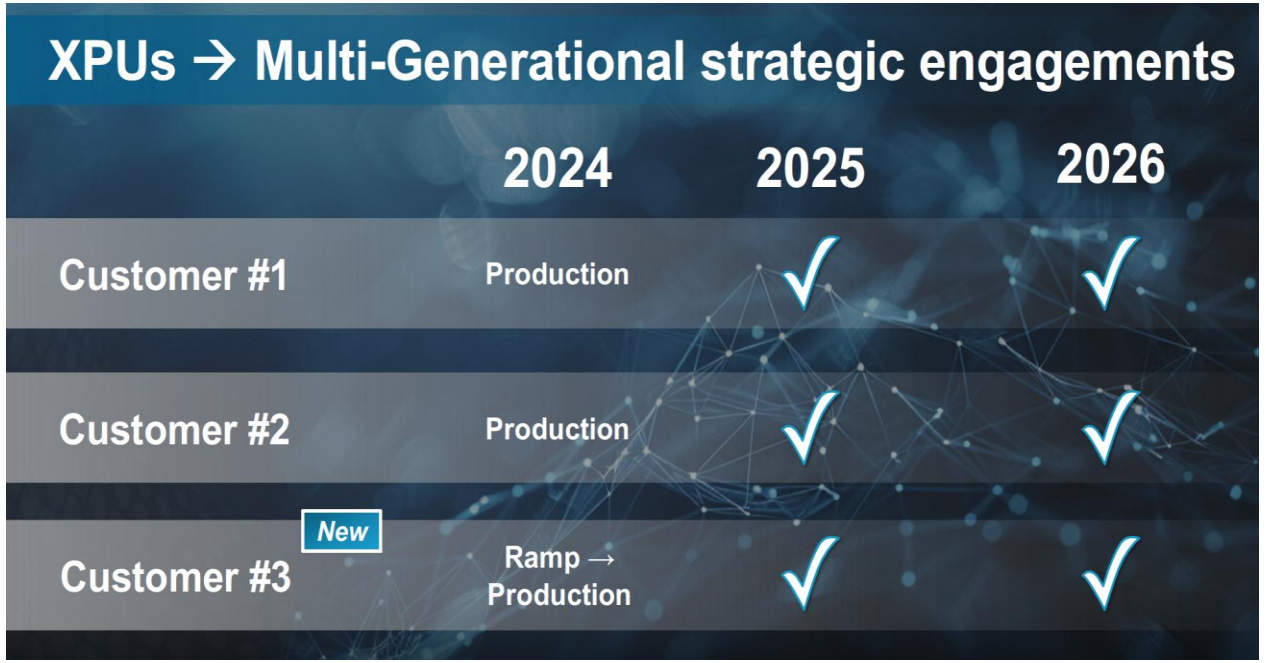
4.1.1 博通AI芯片业务目标

- 博通AI业务占比从2019年的低于5%提升至2023年的15%左右。根据公司的规划，**预计2024年实现超过100亿美元的收入体量，占公司整体收入比例增长至35%。**
- 目前博通已经为两家头部CSP客户批量供应了ASIC产品。此外，博通另一家客户正在产能爬坡中，预计2025年开始贡献业绩。

2024年AI芯片目标占比35%



博通AI芯片客户导入情况

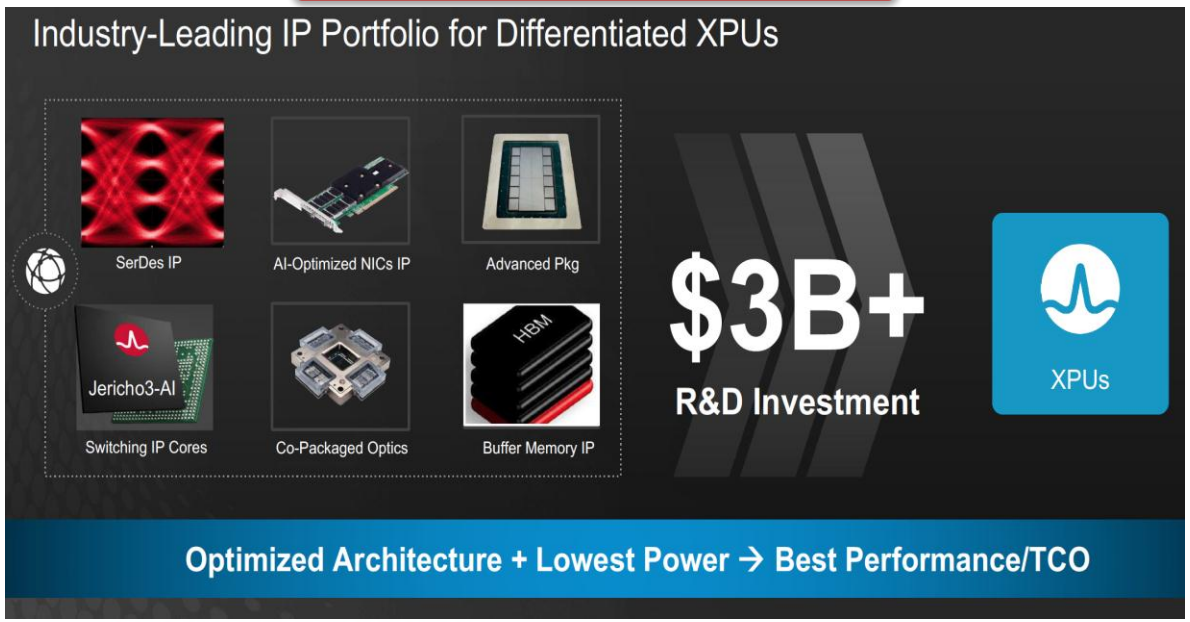


数据来源：博通，西南证券整理

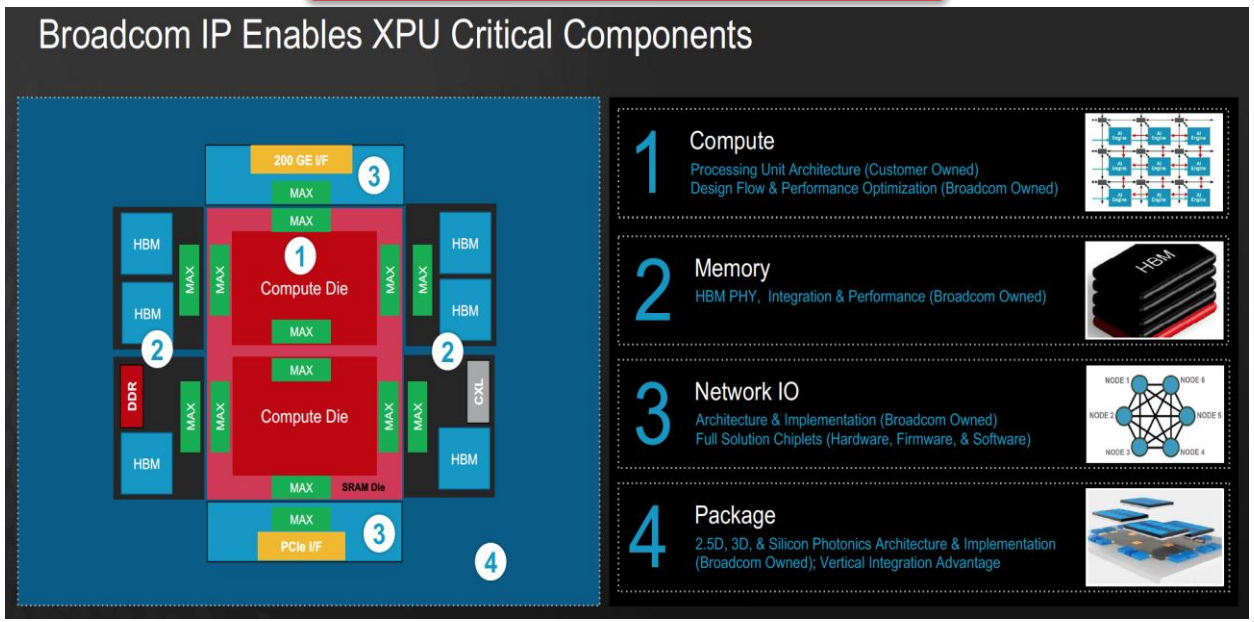
4.1.2 博通广泛的IP储备为ASIC产品线赋能

- 博通广泛的IP储备可为其XPU（ASIC)产品线赋能，博通的IP主要分为4类：计算，存储，网络IO，封装。其中SerDes、基于AI优化NICs、高端封装、交换机、CPO、内存等IP处于行业领先水平。博通在相关领域投入了30亿美元研发费用。
  - ✓ 计算：处理单元架构，设计流和性能优化
  - ✓ 存储：HBM PHY，整合和性能
  - ✓ 网络IO：架构实现，Chiplets软硬一体化解决方案
  - ✓ 封装：2.5D/3D封装，硅光架构和实现，垂直整合等

博通在XPU业界领先的IP能力



博通广泛的IP储备为XPU赋能

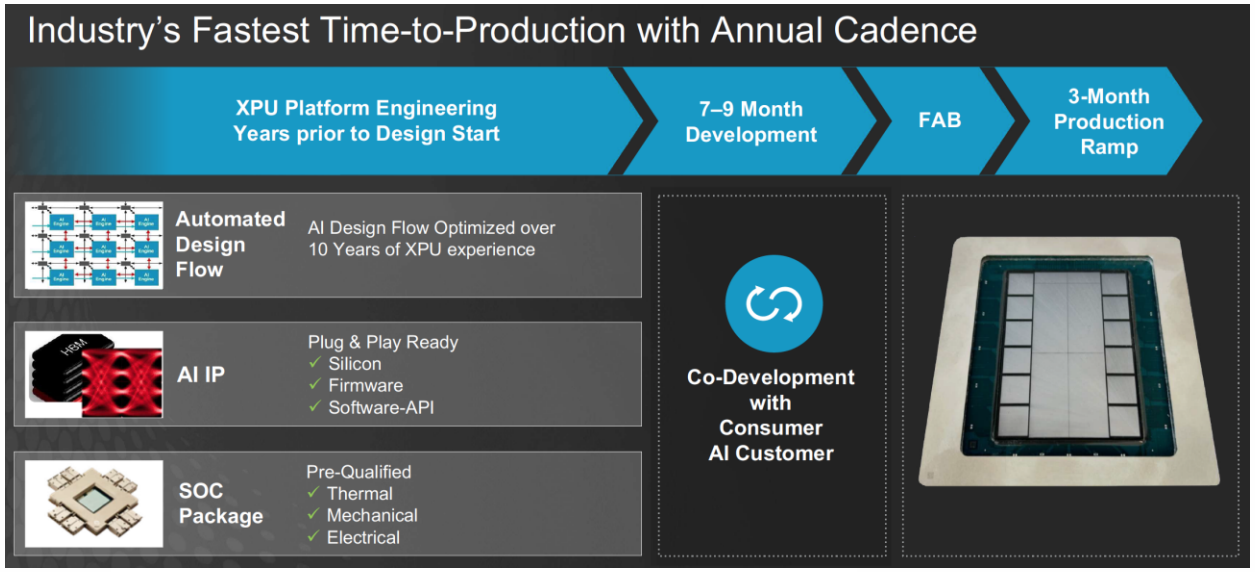


数据来源：博通，西南证券整理

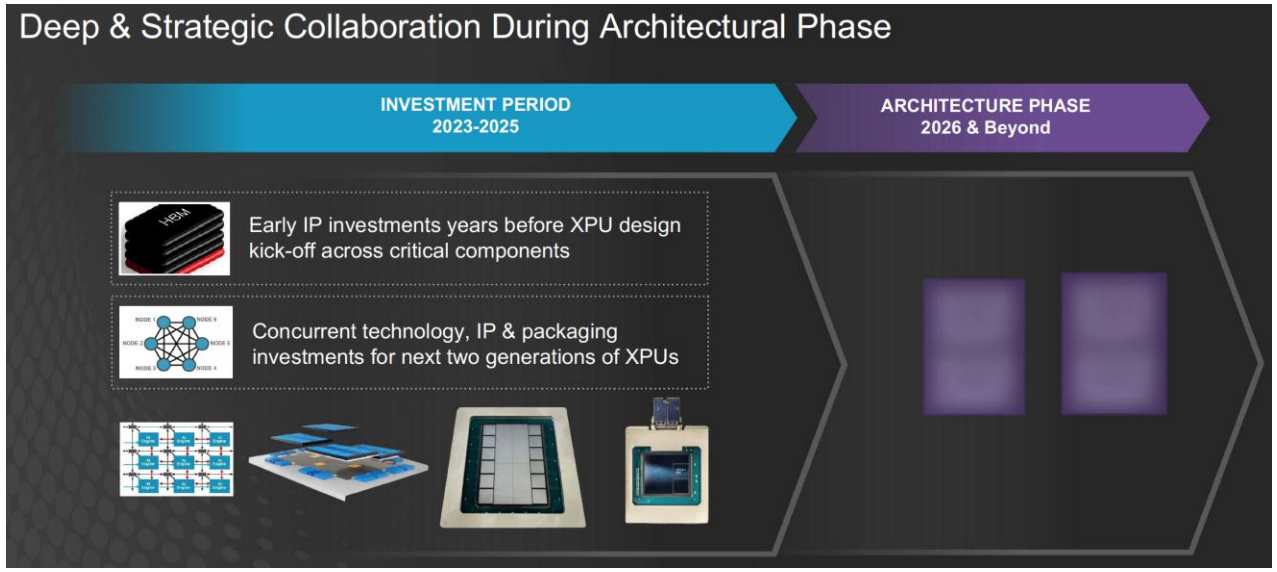
4.1.3 博通利用XPU平台和与客户的深度战略合作实现产品快速落地

- 博通充分利用已经布局完成的XPU平台工程，实现了业界最快的ASIC产品落地时间。XPU平台涵盖了经过10年XPU经验优化的AI设计流程、AI IP、SoC封装等一体化解决方案。ASIC产品设计阶段耗费7-9个月的联合开发时间，再用3个月左右的时间完成产品的生产和产能爬坡。
- 博通与客户在架构阶段就展开了深度的战略合作。在XPU设计启动的前几年，完成了关键组件的早期IP投资；并且为后两代XPU同时进行技术、IP和封装投资。

博通利用XPU平台实现业界最快的产品落地



博通与客户在架构阶段的深度战略合作

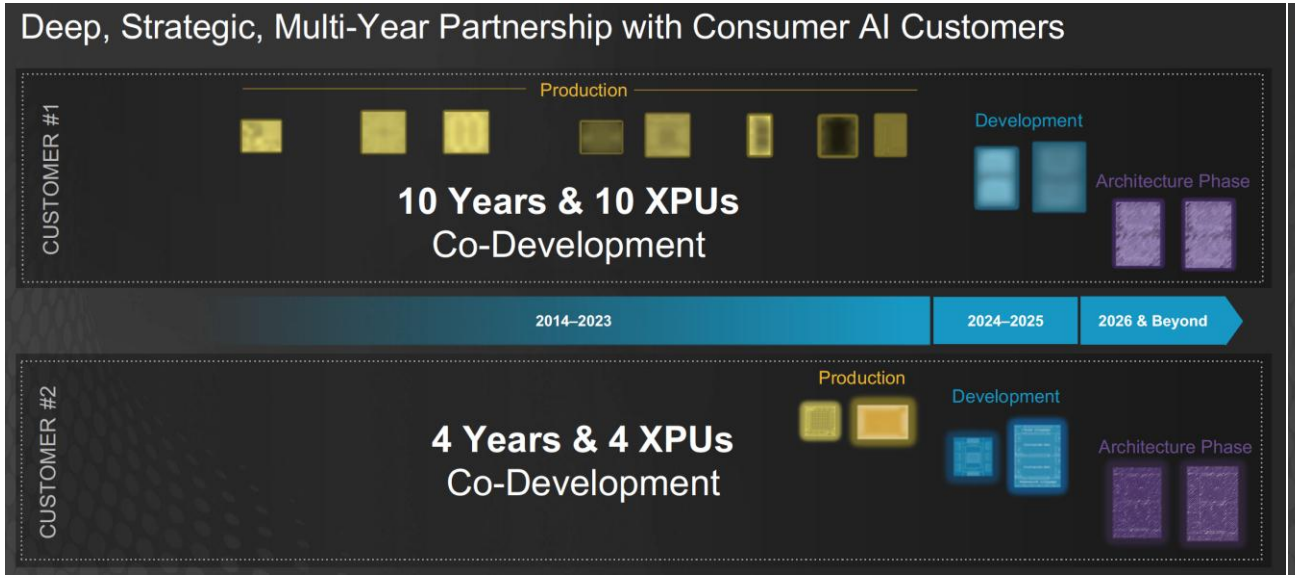


数据来源：博通，西南证券整理

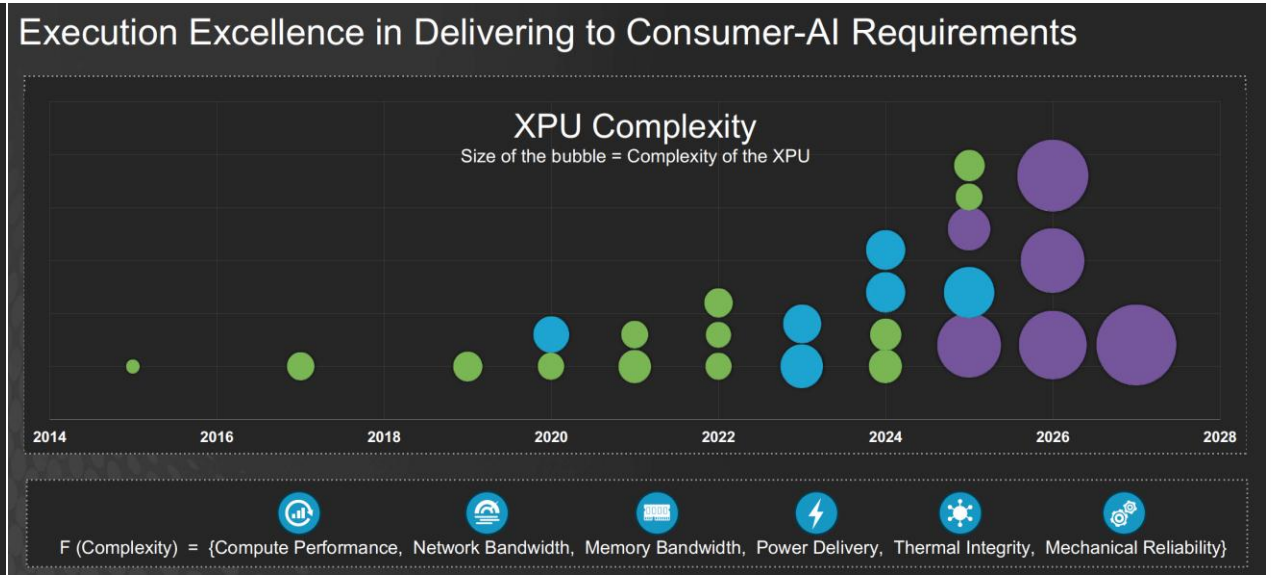
4.1.4 博通与客户联合开发多款不同复杂度的XPU

- 随着算力性能增长、网络和内存带宽提升、对电力输送、热完整度、机械可靠性等要求的升级，XPU的复杂度在加深。博通积极响应了客户对更复杂的XPU的需求。博通与多家大客户一起联合开发了十几款XPU产品。

博通与大客户联合开发了多款XPU



博通满足客户对XPU复杂度的需求

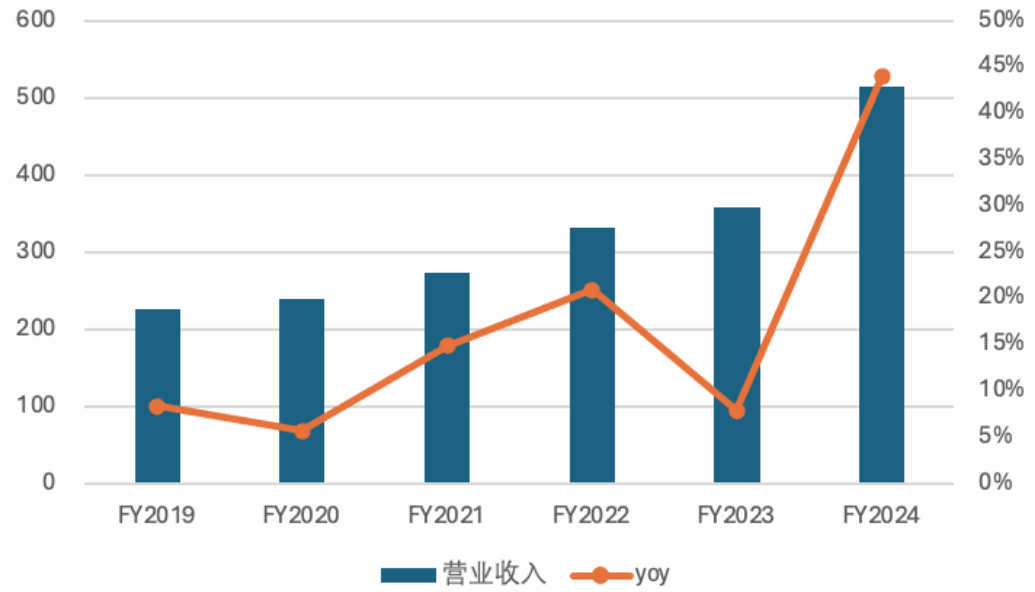


数据来源：博通，西南证券整理

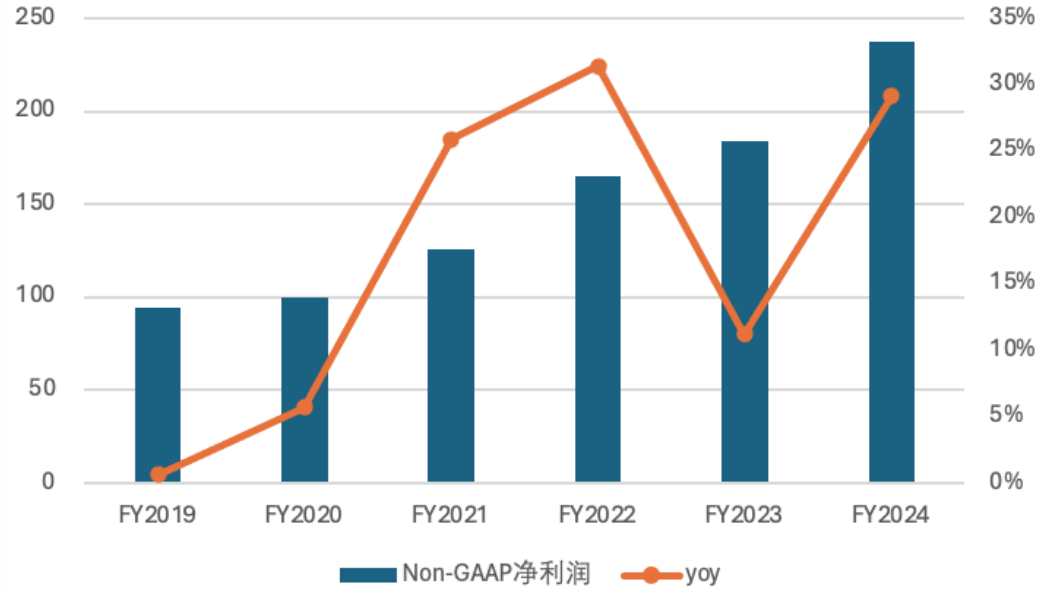
4.1.5 博通收入稳步增长，利润率显著提升

- 博通2019-2023财年收入年复合增速11.4%。2024年公司收购的Vmware并表后，2024财年实现收入515.7亿美元，同比增长44%。
- 博通Non-GAAP净利率从2019财年的41.8%提升至2023财年的51.3%。得益于利润率的提升，博通2019-2023财年Non-GAAP净利润年复合增速达14.4%。2024财年Non-GAAP净利润达到237.3亿美元，同比增长29%。

博通收入（亿美元）



博通净利润（亿美元）

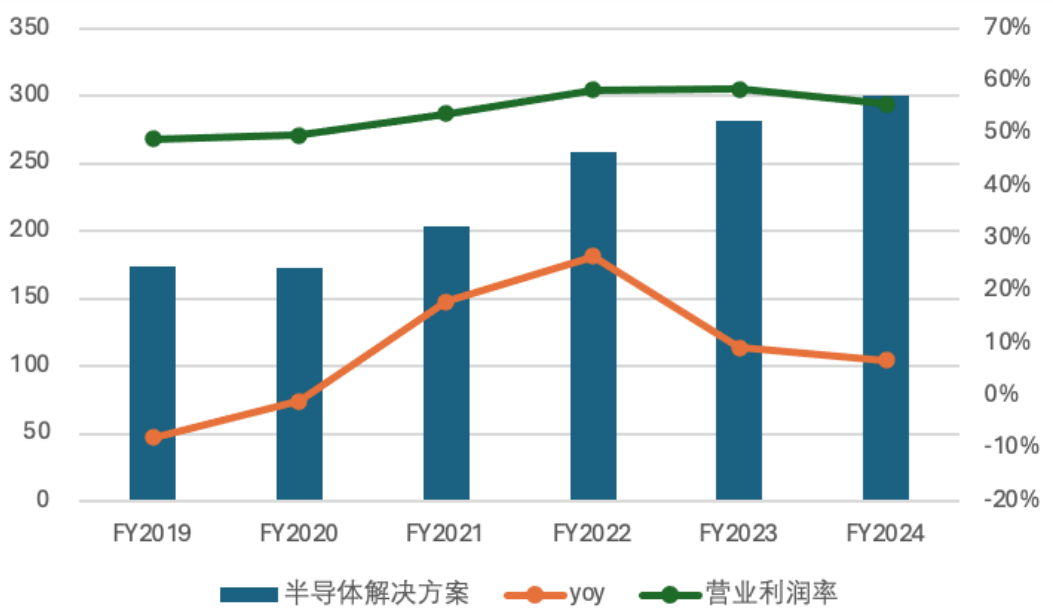


数据来源：博通公告，西南证券整理

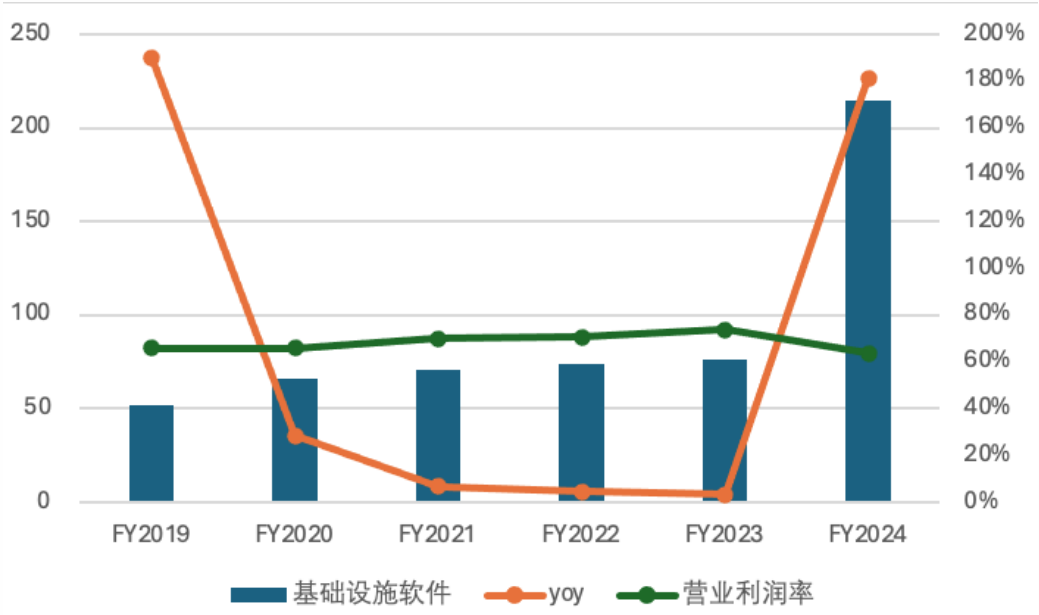
4.1.6 半导体是博通最大业务，盈利能力逐年提升

- 半导体解决方案是公司的最大业务，2023财年占据公司整体收入的78.7%；利润率从2019财年的50%提升至2023年的58.5%。2024财年公司半导体业务收入301亿美元，同比增长6.8%。
- 公司的基础设施软件业务占比不低于20%。其中，Vmware在2024年并表后显著推高了该业务板块的规模体量。

博通半导体业务收入（亿美元）



博通基础设施软件收入（亿美元）

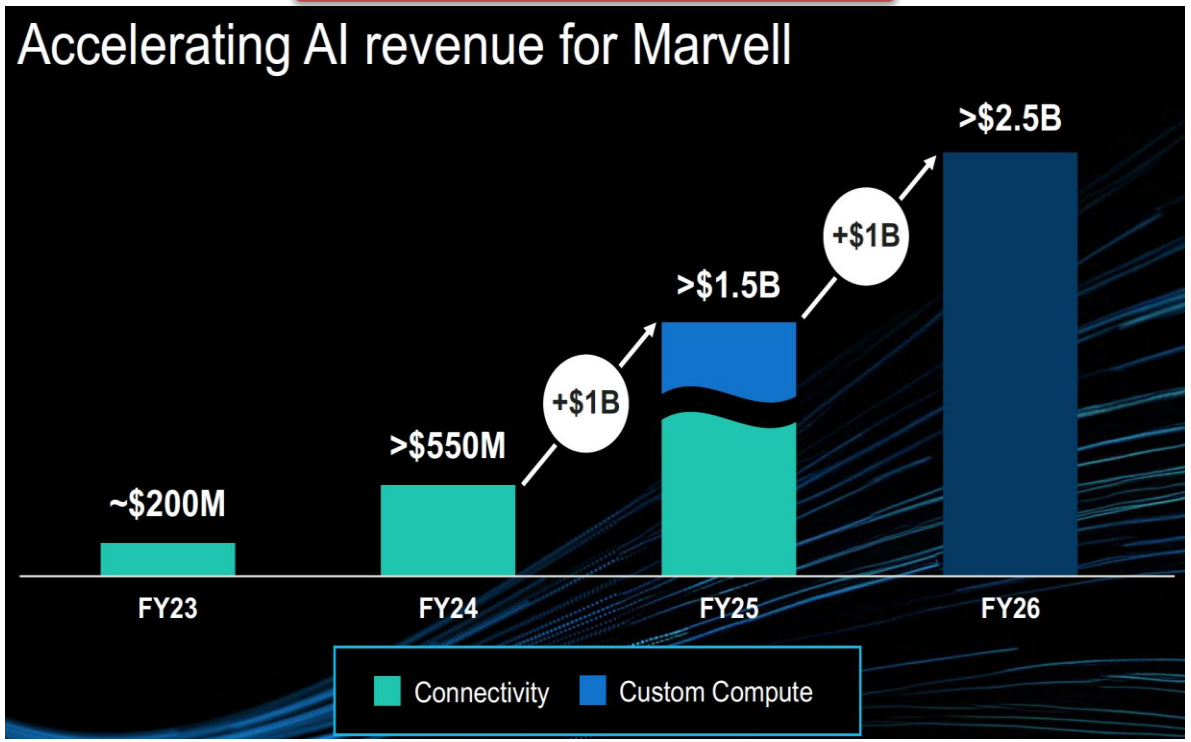


数据来源：博通公告，西南证券整理

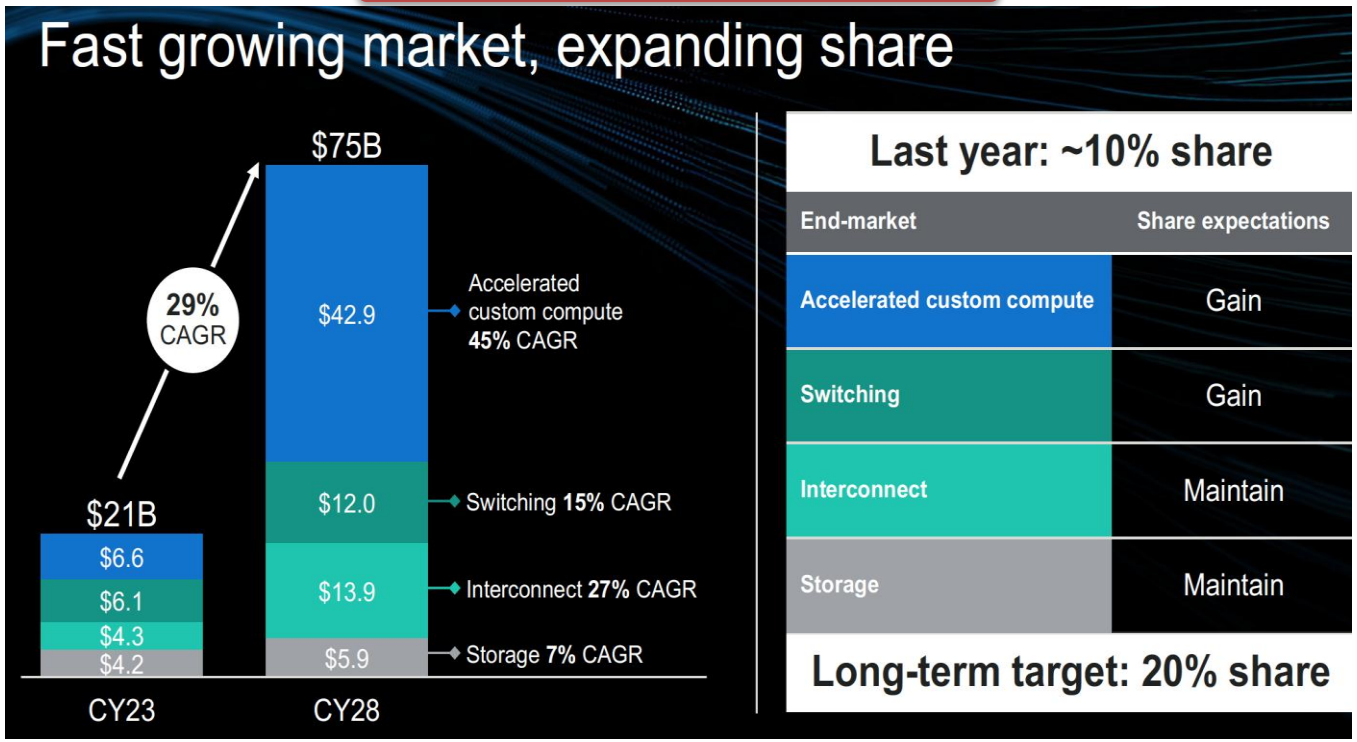
4.2.2 Marvell AI业务目标

- Marvell的AI业务2023财年为2亿美元左右。**公司预计24-26财年加速AI业务（连接+定制化计算）收入从5.5亿提升至25亿美元。**
- Marvell的数据中心业务TAM：根据Marvell预测，**2023-2028年其数据中心业务TAM从210亿美元增长至750亿美元，CAGR为29%**；其中，**定制化加速计算TAM从66亿美元增长至429亿美元，CAGR为45%**；交换机TAM从61亿美元增长至120亿美元，CAGR为15%；互联TAM从43亿美元增长至139亿美元，CAGR为27%；存储市场从42亿美元增长至59亿美元，CAGR为7%。
- **Marvell数据中心业务23年市占率10%，公司长期市占率目标为20%，即业务规模150亿美元，相当于23-28年CAGR高达46.6%。**

Marvell加速AI业务收入规模预测



Marvell数据中心TAM

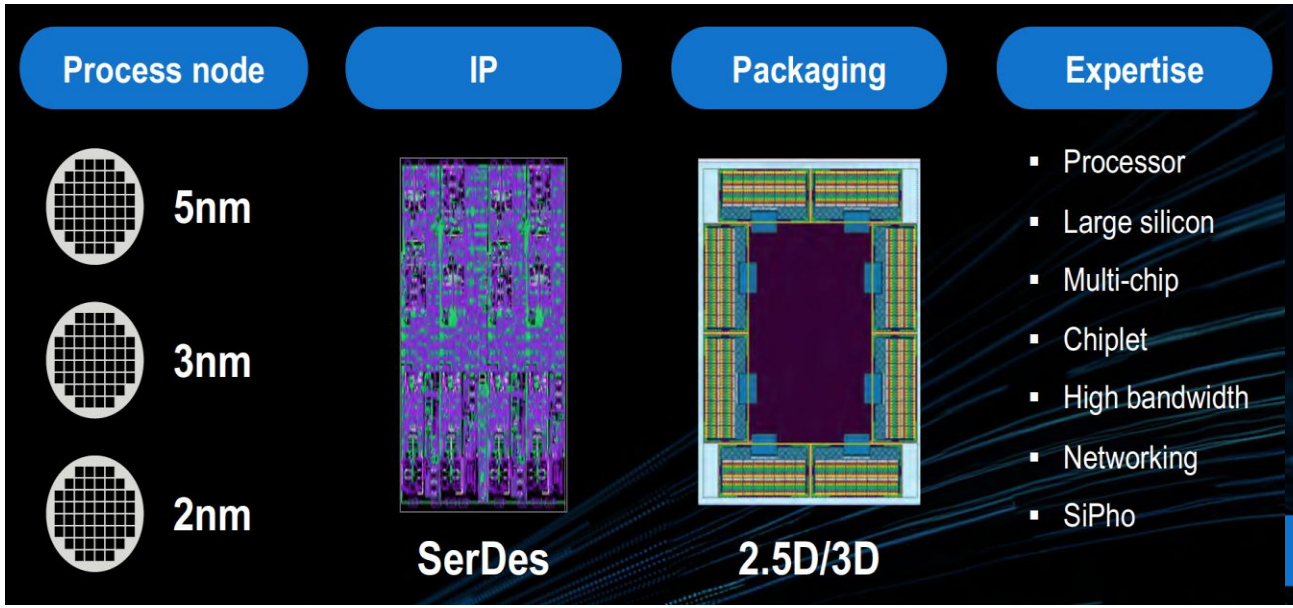


数据来源：Marvell，650 Group, CignalAI, Dell'Oro, LightCounting，西南证券整理

4.2.3 Marvell的加速计算基础设施平台布局

- Marvell的加速计算基础设施平台涵盖了“工艺制程-IP-封装-专家”的布局。
- Marvell经过多年对完整定制平台的收购和重大投资，储备了一大批世界级的IP，覆盖模拟、数字、封装等多个层面的知识产权。其中，Cavium擅长网络加速计算，AveraSemi（原格芯子公司）擅长为各种应用提供定制芯片解决方案和2.5D/3D封装技术，Aquantia擅长网络传输，Inphi擅长模拟、硅光和DSP技术，Innovium擅长数据中心交换机芯片技术（竞品为博通的Trident和Tomahawk芯片）。

Marvell的加速计算基础设施平台



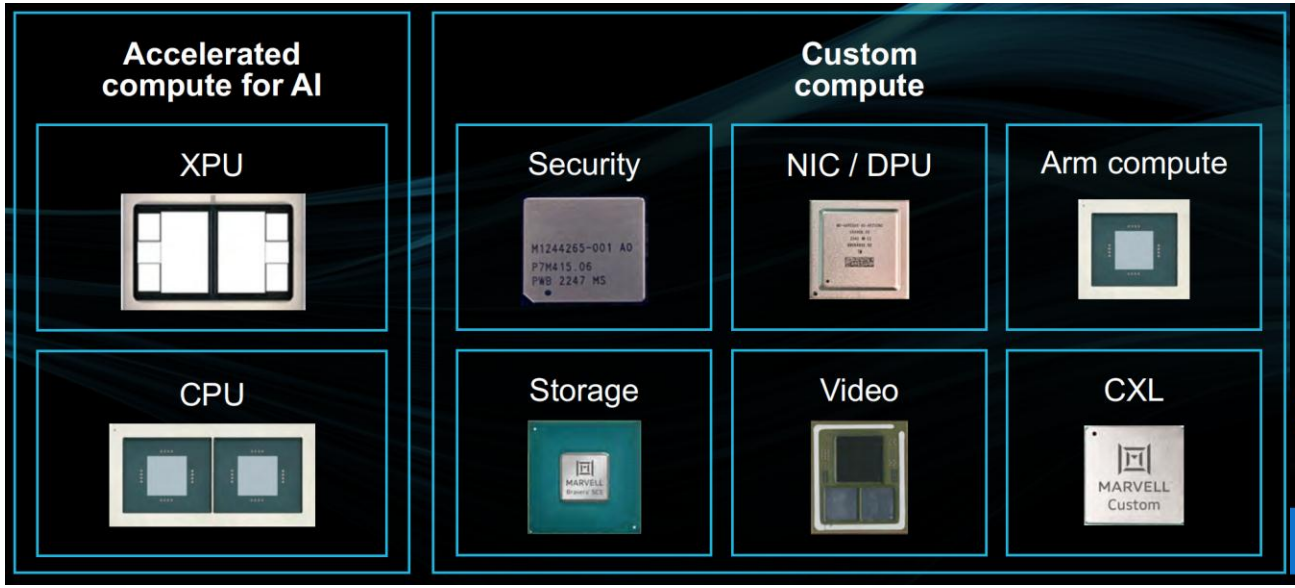
Marvell的IP体系

数据来源：Marvell，西南证券整理

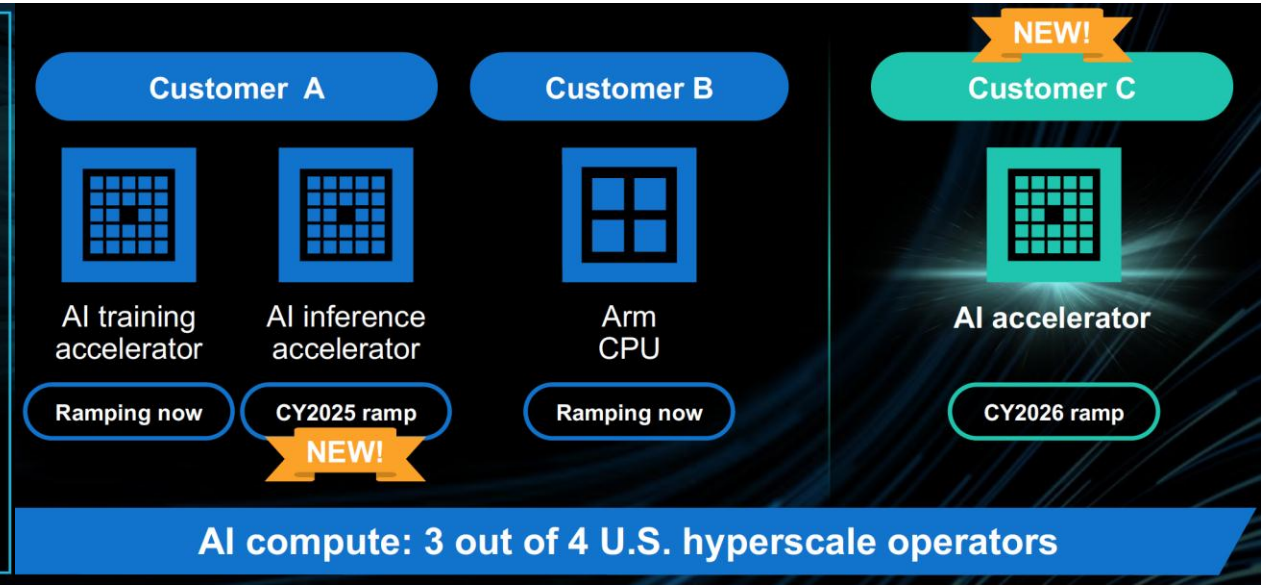
4.2.4 Marvell的定制计算产品线和主要客户情况

- Marvell的定制计算产品包括AI加速芯片，针对安全、NIC/DPU、ARM计算、存储、视频和CXL功能的ASIC等。
- Marvell的客户涵盖美国3/4的大型CSP。Marvell为亚马逊设计的AI训练加速器Trainium 2已批量出货。B客户的ARM CPU正处于产能爬坡阶段。新介入的C客户其AI加速器将于2026年产能爬坡。

Marvell多样化的定制计算产品线



Marvell AI产品的主要客户

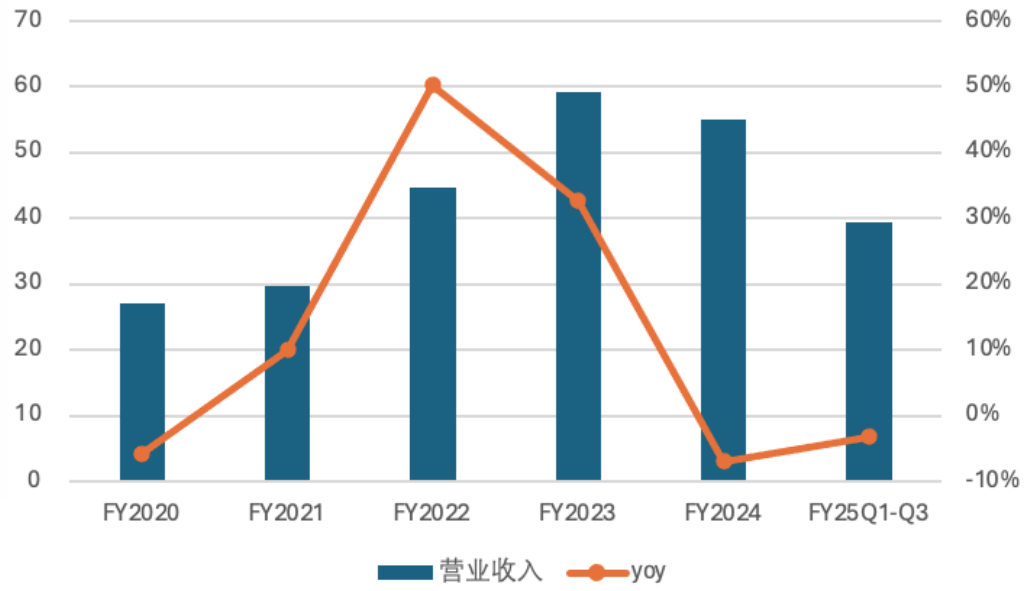


数据来源：Marvell，西南证券整理

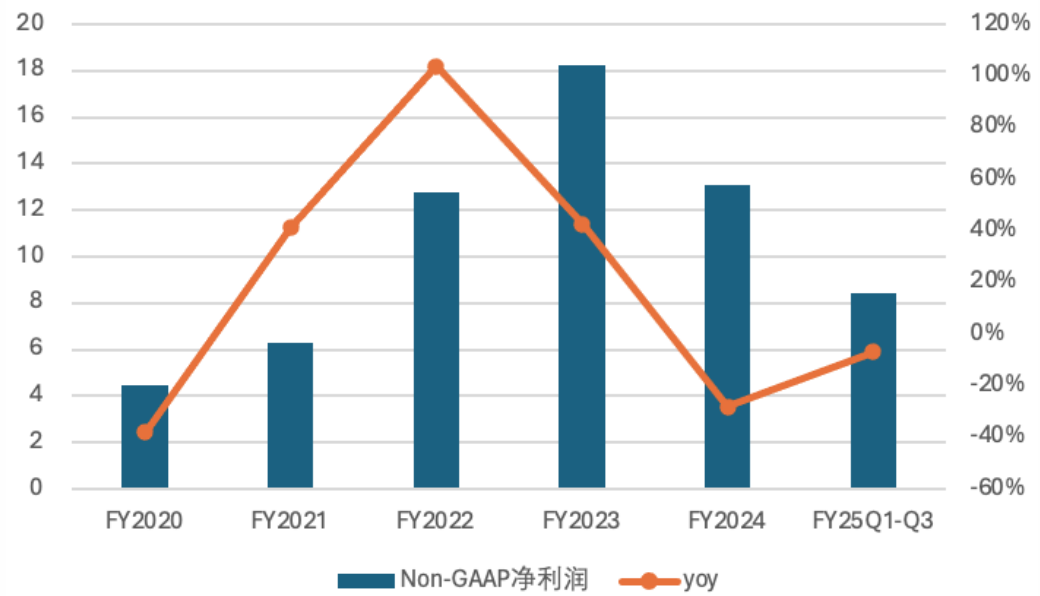
4.2.5 Marvell业绩弹性大，2025财年盈利逐季改善

- Marvell 2021-2024财年收入年复合增速19.5%。2025财年随着经营改善，2025前三财季收入逐季改善（同比增速依次为-12.2%、-5%、6.9%）；FY2025Q3单季度收入15.2亿美元，同比增长6.9%，增速重新转正。
- Marvell 2021-2024财年Non-GAAP净利润年复合增速达31%。2025前三财季Non-GAAP净利润为8.5亿美元，Non-GAAP净利率从17.8%提升至24.6%。

Marvell收入（亿美元）



Marvell净利润（亿美元）

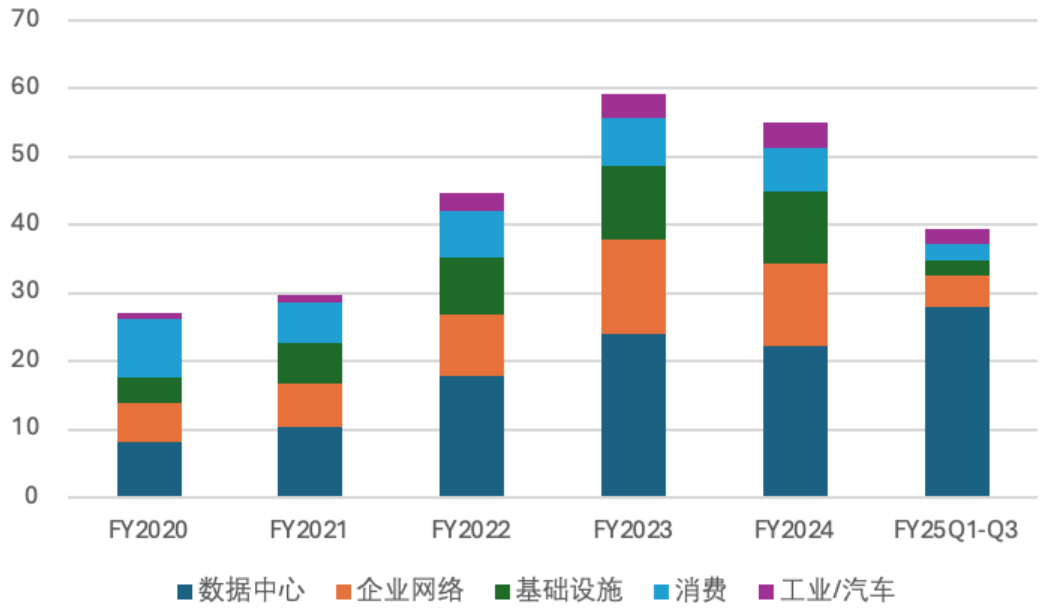


数据来源：Marvell公告，西南证券整理

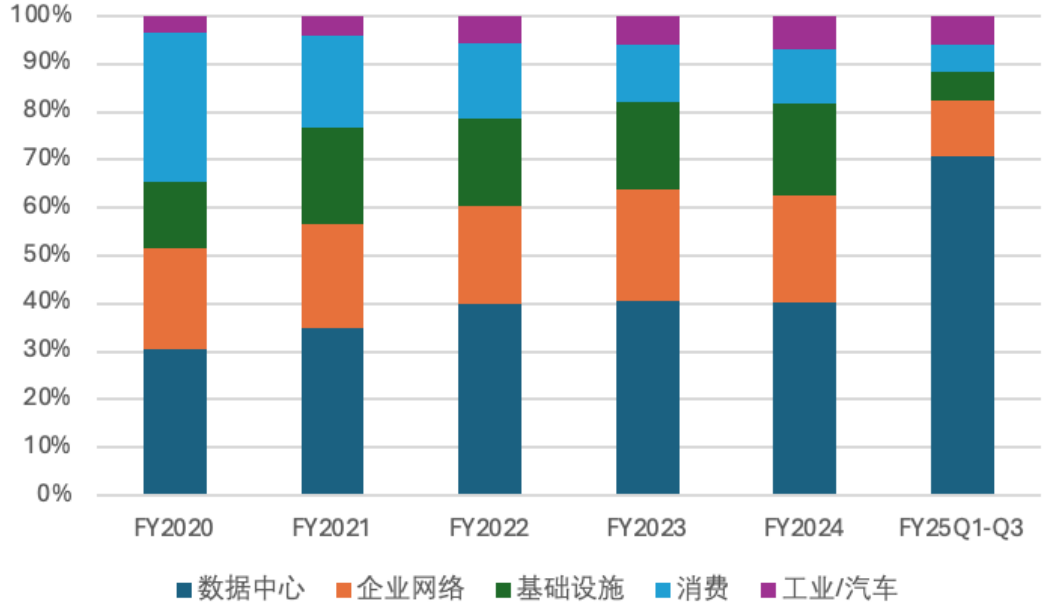
4.2.6 受益于AI需求推动，数据中心业务比重超40%

- 数据中心是公司的最大业务，2021-2024财年复合增速28%，高于整体收入增速。受益于AI对光学等产品的需求推动，最近两个财年数据中心业务占据整体收入比重超40%。
- 其他业务中，企业网络收入2021-2024财年复合增速21.2%，2024财年占收入比重22%左右；基础设施收入2021-2024财年复合增速23.2%，2024财年占收入比重19%左右；消费业务2024财年占收入比重11%左右；工业和汽车业务2021-2024财年复合增速44%，2024财年占收入比重7%左右。

Marvell分业务收入（亿美元）



Marvell分业务收入占比



数据来源：Marvell公告，西南证券整理

风险提示

- ✓ AI产业发展不及预期的风险；
- ✓ 互联网科技企业资本支出不及预期的风险；
- ✓ GPU竞争的风险。

分析师：王湘杰
执业证号： S1250521120002
电话：0755-26671517
邮箱：wxj@swsc.com.cn

分析师：杨镇宇
执业证号：S1250517090003
电话：023-67563924
邮箱：yzyu@swsc.com.cn

西南证券投资评级说明

报告中投资建议所涉及的评级分为公司评级和行业评级（另有说明的除外）。评级标准为报告发布日后6个月内的相对市场表现，即：以报告发布日后6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。其中：A股市场以沪深300指数为基准，新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准；美国市场以纳斯达克综合指数或标普500指数为基准。

公司评级	买入：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在20%以上
	持有：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于10%与20%之间
	中性：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-10%与10%之间
	回避：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-20%与-10%之间
	卖出：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在-20%以下
行业评级	强于大市：未来6个月内，行业整体回报高于同期相关证券市场代表性指数5%以上
	跟随大市：未来6个月内，行业整体回报介于同期相关证券市场代表性指数-5%与5%之间
	弱于大市：未来6个月内，行业整体回报低于同期相关证券市场代表性指数-5%以下

分析师承诺

报告署名分析师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，报告所采用的数据均来自合法合规渠道，分析逻辑基于分析师的职业理解，通过合理判断得出结论，独立、客观地出具本报告。分析师承诺不曾因，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获取任何形式的补偿。

重要声明

西南证券 SOUTHWEST SECURITIES

西南证券研究发展中心

西南证券研究发展中心

上海

地址：上海市浦东新区陆家嘴21世纪大厦10楼

邮编：200120

北京

地址：北京市西城区金融大街35号国际企业大厦A座8楼

邮编：100033

深圳

地址：深圳市福田区益田路6001号太平金融大厦22楼

邮编：518038

重庆

地址：重庆市江北区金沙门路32号西南证券总部大楼21楼

邮编：400025

西南证券机构销售团队

区域	姓名	职务	手机	邮箱	姓名	职务	手机	邮箱
上海	蒋诗烽	总经理助理/销售总监	18621310081	jsf@swsc.com.cn	欧若诗	销售经理	18223769969	ors@swsc.com.cn
	崔露文	销售副总监	15642960315	clw@swsc.com.cn	李嘉隆	销售经理	15800507223	ljlong@swsc.com.cn
	李煜	高级销售经理	18801732511	yfliyu@swsc.com.cn	龚怡芸	销售经理	13524211935	gongyy@swsc.com.cn
	田婧雯	高级销售经理	18817337408	tjw@swsc.com.cn	孙启迪	销售经理	19946297109	sqdi@swsc.com.cn
	张玉梅	销售经理	18957157330	zymyf@swsc.com.cn	蒋宇洁	销售经理	15905851569	jyj@swsc.com.c
	魏晓阳	销售经理	15026480118	wxyang@swsc.com.cn				
北京	李杨	销售总监	18601139362	yfly@swsc.com.cn	张鑫	高级销售经理	15981953220	zhxin@swsc.com.cn
	张岚	销售副总监	18601241803	zhanglan@swsc.com.cn	王一菲	高级销售经理	18040060359	wyf@swsc.com.cn
	杨薇	资深销售经理	15652285702	yangwei@swsc.com.cn	王宇飞	高级销售经理	18500981866	wangyuf@swsc.com
	姚航	高级销售经理	15652026677	yhang@swsc.com.cn	马冰竹	销售经理	13126590325	mbz@swsc.com.cn
广深	郑龑	广深销售负责人	18825189744	zhengyan@swsc.com.cn	杨举	销售经理	13668255142	yangju@swsc.com.cn
	杨新意	广深销售联席负责人	17628609919	yxy@swsc.com.cn	陈韵然	销售经理	18208801355	cyryf@swsc.com.cn
	龚之涵	高级销售经理	15808001926	gongzh@swsc.com.cn	林哲睿	销售经理	15602268757	lzr@swsc.com.cn
	丁凡	销售经理	15559989681	dingfyf@swsc.com.cn				