GPGPU 与 ASIC 之争——算力芯片看点系列

2025 年 03 月 12 日

增持（维持）

投资要点

- **GPGPU 与 ASIC 性能对比一览** 我们梳理各芯片参数，得到如下结论：**1）算力方面，**多数 ASIC 较少涉及高精度浮点数数据，聚焦于低精度领域且拥有相对而言更可观的功耗控制与能效比，但尽管在低精度领域，算力性能部分指标仍难以与同时期的 GPGPU 相媲美。**2）存力方面，**ASIC 算力密度高，算数强度迭代快，但在显存带宽和容量上与 GPGPU 仍有较大差距，近期表现亮眼的 LPU 则通过超高内存带宽突破性化解传统 GPU 的内存瓶颈。**3）互连方面，**英伟达 NVLink 所能实现的 Scale-up 互连能力一骑绝尘，挑战英伟达 NVLink 的难度较大。ASIC 在特定性能上表现突出，但整体来看仍较难超越英伟达的市场地位。

- **为什么大厂纷纷开始自研 AI 芯片？** 芯片公司的支出通常包含员工薪资、EDA 和 IP 费用、芯片制造费用、销售费用四个方面，我们按 Fabless 公司的研发投入模式，依据寒武纪、海光信息、翱捷科技的研发人员人数与薪酬数据及英伟达相关产品售价与销售毛利进行计算，大约 4.5-7 万卡出货量可以覆盖前期的投入。而头部大厂的万卡集群建设未曾停歇，完全有望覆盖自研 ASIC 的前期投入，训练端单一集群的需求量已逐渐超过 10 万卡，同时英伟达 FY2024 数据中心有 40%的收入来自推理业务。随着 AI 应用遍地开花，我们认为 AI 推理需求还有更大渗透空间。

- **大厂自研 AI 芯片谁能代工？** 博通产品线 IP 生态强大完善，在接口、互连等领域保持前瞻性优势，针对不同规模的 AI 集群提供差异化系统架构与解决方案，并在 2024 年发布了业界首款采用 5nm CMOS 工艺实现的 400 千兆以太网（GbE）NIC 设备第二代网卡芯片 Thor 2。Marvell 通过 HBM 重构与 CPO 集成的双重突破体现竞争力，直击 AI 芯片的能效与带宽瓶颈，AI 业务增长显著。台企世芯电子（AIchip）、创意电子（GUC）展现出先进制程与系统设计的优势。中兴通讯作为全球知名的通信与信息技术解决方案提供商，在算力基础设施领域掌握多项核心技术，构建起完备的技术体系。翱捷科技具备完备齐全的自研 IP，在蜂窝基带芯片、非蜂窝物联网芯片设计方面经验丰富，同时拥有成熟的芯片定制服务能力。芯原股份为客户提供平台化、全方位、一站式芯片定制服务和半导体 IP 授权服务，为国内大厂自研提供更多的代工选择。

- **投资建议：**推荐寒武纪、海光信息（与计算机组共同覆盖），建议关注中兴通讯、翱捷科技、芯原股份。

- **风险提示：**大厂 CapEx 投入不及预期，技术发展不及预期，客户需求不及预期。

行业走势

相关研究

《国产算力腾飞，看好 Ascend 910C 产业链》

2025-03-05

表 1：重点公司估值

代码	公司	总市值（亿元）	收盘价（元）	EPS			PE			投资评级
				2023A	2024E	2025E	2023A	2024E	2025E	
688256	寒武纪-U	3,306.26	792.00	-2.03	-1.10	0.51	-389.69	-720.00	1,552.94	买入
688041	海光信息	3,672.69	158.01	0.54	0.83	1.25	290.75	190.20	126.41	买入

数据来源：Wind，东吴证券研究所预测

注：海光信息 2024 年报已发布，此处 2024 数据为历史数据；收盘价截至 2025/3/11

内容目录

图表目录

1. GPGPU 与 ASIC 性能对比一览

1.1. 算力：精度与能效的差异化竞争

1）从精度范围来看，ASIC 较少涉及高精度浮点数数据，主要聚焦于低精度领域，这与其主要应用于大模型训练的定位相符。大模型训练过程中，低精度数据类型（如 INT8、FP16 等）足以满足大部分计算需求，并且能够在一定程度上减少计算量和存储需求，提高训练效率。**2）就低精度部分的算力性能而言，大厂自研的 ASIC 在一些指标上也难以与同时期的 GPGPU 相媲美。**以英伟达 GB200 为例，FP16 达 5000，远超同时期 ASIC 数值。**3）在功耗和能效比方面，多数 ASIC 拥有相对而言更可观的功耗控制与能效比。**通常，ASIC 由于其定制化的设计，专为特定任务（如大模型训练）优化，在执行特定任务时可能具有相对较低的功耗。GPGPU 在执行相同任务时，由于其架构需要兼顾多种计算场景，功耗往往较高。例如，微软的 Maia 100 能效比高达 1.60，而同时期的英伟达 H200 为 1.41。但也有例外，如英伟达 A100 的能效比（0.78）高于同期谷歌 TPU v4i（0.39），呈现出兼顾普适性与高效性的特点。

图1：主流 AI 芯片算力指标梳理

厂商	名称	发布时间	FP64	FP64矩阵	FP32	FP32矩阵	TF32矩阵	BF16/FP16/FP16矩阵	INT8	FP8	FP6	INT4	FP4	功耗	能效比(FP16)
海外芯片															
英伟达	GB200	2024/3/19	90	90	180		2500	5000	10000	10000	10000		20000	2700	1.85
	H200	2023/11/13	34	67	67		494.5	989	1979	1979				700	1.41
	H100	2022/3/22	34	67	67		494.5	989	1979	1979				700	1.41
	H800	2023/3/22	1	1	67		494.5	989	1979	1979				700	1.41
	A100	2020/5/14	9.7	19.5	19.5		156	312	624					400	0.78
	A800	2022/11/8	9.7	19.5	19.5		156	312	624					400	0.78
	V100	2017/5/11	7.8		15.7			125	62					300	0.42
AMD	MI350X	2024/6/3													
	MI325X	2024/6/3													
	MI300X	2023/12/6	81.7	163.4	163.4	163.4	653.7	1307.4	2614.9	2614.9				750	1.74
	MI300A	2023/12/6	61.3	122.6	122.6	122.6	490.3	980.6	1961.2	1961.2				550 | 760	1.78 | 1.29
	MI250X	2021/11/8	47.9	95.7	47.9	95.7		383	383			383		500 | 560	0.76 | 0.68
	MI250	2021/11/8	45.3	90.5	45.3	90.5		362.1	362.1			362.1		500 | 560	0.72 | 0.65
	MI210	2022/3/22	22.6	45.3	22.6	45.3		181	181			181		300	0.60
英特尔	Gaudi3	2024/4/9			14.3	229	459	1835		1835				900	2.04
	Gaudi2	2022/5/1												600	0.00
Groq	LPU	2020/9/23						188	750					275	0.68
海外大厂自研															
谷歌	TPU v6e	2024/5/15						926	1852					600	1.54
	TPU v5p	2023/12/7						459	918						
	TPU v5e	2023/8/30						197	394						
	TPU v4	21Q4						137.5	275					192	0.72
	TPU v4i	20Q1						69	138					175	0.39
亚马逊	Trainium3	2024/12/3						1334	2598					714	1.87
	Trainium2	2023/11/28						667	1299					500	1.334
Meta	MTIA v2	2024/4/10			2.76	2.76		177	708					90	1.97
	MTIA v1	2023/5/18			0.8	1.6		51.2	102.4					25	2.05
微软	Maia 100	2023/11/16						800	1600				3200	500	1.60
特斯拉	D1	2021/8/20	不支持	不支持	22.6	22.6		362	400					400	0.91

数据来源：各公司官网，A5 图王，量子位，新智元，AI 时代前沿，IT 之家，SemiAnalysis，半导体行业观察，智东西，芯智讯，机器之心，芯榜，科闻社，快科技，半导体产业纵横，电厂，电子工程专辑，东吴证券研究所

1.2. 存力：显存性能与算力密度的权衡角逐

1）从显存性能来看，自研 ASIC 在显存带宽和容量上与 GPGPU 仍有较大差距。

GB200 依靠 HBM3e 技术拥有高达 16384GB/s 的带宽，这使其在处理大规模数据时能更高效地运行复杂任务。**2）从算力密度（算力/显存容量）来看，GPGPU 单位显存算力相对有限，ASIC 则以高算力密度在特定任务凸显优势。**在实际应用中，较高的算力密度意味着在相同的显存资源下，芯片能够完成更多的计算任务。以谷歌 TPU v6e 为例，FP16 算力 1852，显存容量 32GB，算力密度约 57.88，展现出显存利用效率高、存力与算力协同性好的特征。**3）从算术强度（算力/显存带宽）来看，早期 ASIC 弱于同时期 GPU，但技术迭代速度快，22 年后实现反超。**至 24 年，ASIC 芯片如 Meta MTIA v2 算术强度达 885 FLOPs/Byte，是同期 GB200 算术强度的 2.8 倍。**4）LPU 通过超高内存带宽突破性化解传统 GPU 的内存瓶颈。**LPU 采用 230MB SRAM 集成设计，提供 80TB/s 的峰值内存带宽。这种存力使每个计算单元可即时获取连续 token 序列，消除传统架构中因频繁访问外部显存产生的时钟周期损耗。该设计架构通过存力创造性释放算力潜能，为大模型推理提供数据供给保障，完成低算术强度任务性能创造性突破。

图2：主流 AI 芯片存力指标梳理

厂商	名称	发布时间	HBM/显存	显存带宽（GB/s）	显存容量（GB）	算术强度（FP16）
海外芯片						
英伟达	GB200	2024/3/19	HBM3e	16384	384	312.5
	H200	2023/11/13	HBM3e	4915.2	141	206.0
	H100	2022/3/22	HBM3	3430.4	80	295.2
	H800	2023/3/22	HBM3	3430.4	80	295.2
	A100	2020/5/14	HBM2e	2039	80	156.7
	A800	2022/11/8	HBM2e	2039	80	156.7
	V100	2017/5/11	HBM2	900	32	142.2
AMD	MI350X	2024/6/3	HBM3e		288	
	MI325X	2024/6/3	HBM3e	6144	288	
	MI300X	2023/12/6	HBM3	5427.2	192	246.7
	MI300A	2023/12/6	HBM3	5427.2	128	185.0
	MI250X	2021/11/8	HBM2e	3276.8	128	119.7
	MI250	2021/11/8	HBM2e	3276.8	128	113.2
	MI210	2022/3/22	HBM2e	1638.4	64	113.1
英特尔	Gaudi3	2024/4/9	HBM2e	3788.8	128	495.9
	Gaudi2	2022/5/1	HBM2e	2519.04	96	0.0
Groq	LPU	2020/9/23	无，使用SRAM	81920	0.22	2.4
海外大厂自研						
谷歌	TPU v6e	2024/5/15	HBM3或3e	1640	32	559.8
	TPU v5p	2023/12/7	HBM3	2765	95	170.0
	TPU v5e	2023/8/30	HBM2	819	16	246.3
	TPU v4	21Q4	HBM2	1228	32	114.7
	TPU v4i	20Q1		300	8	235.5
亚马逊	Trainium3	2024/12/3				230.0
	Trainium2	2023/11/28	HBM3	2900	96	230.0
Meta	MTIA v2	2024/4/10	LPDDR5	204.8	128	885.0
	MTIA v1	2023/5/18	LPDDR5	176	64	297.9
微软	Maia 100	2023/11/16	HBM2e	1843.2	64	444.4
特斯拉	D1	2021/8/20	HBM	800	32	463.4

数据来源：各公司官网，A5 图王，量子位，新智元，AI 时代前沿，IT 之家，SemiAnalysis，半导体行业观察，智东西，芯智讯，机器之心，芯榜，科闻社，快科技，半导体产业纵横，电厂，电子工程专辑，东吴证券研究所

注：算术强度（Arithmetic Intensity），即计算操作数（FLOPs）与内存访问量（Bytes）的比值，用于衡量计算任务的性能瓶颈。高算术强度的任务受限于计算能力，低算术强度的任务受限于内存带宽。

1.3. 互连：NVLink 主导下的技术挑战与突破

1）单从纸面性能来看，英伟达 NVLink 所能实现的 Scale-up 互连能力一骑绝尘。GB200 所依赖的 NVLink5.0 技术能够实现 1.8TB/s 的互连速度，而其他厂商的 Scale-up 互连大多以 PCIe 协议为基础，目前 PCIe5.0 技术单通道双向速率为 8GB/s，16 通道可达 128GB/s，远远低于 NVLink 同代技术。**2）从技术节奏来看，挑战英伟达 NVLink 的难度较大。**UALink 初代 V1.0 标准将于 25Q1 发布，NVLink1.0 早在 2016 年已应用于 Pascal 架构 GPU。

图3：主流 AI 芯片互连指标梳理

厂商	名称	发布时间	Scale-up互连技术	互连速度（GB/s）	底层协议	协议速率（GB/s）	单通道速率（Gbps）
第三方芯片							
英伟达	GB200	2024	NVLink5	2×1800	NVLink5	1800	200
	H200	2023	NVLink4	900	NVLink4	900	
	H100	2023		900			
	H800	2023	NVLink	400	/	/	
	A100	2020	NVLink3	600	NVLink3	600	
	A800	2022	NVLink	400	/	/	
	V100	2017	NVLink2	300	NVLink2	300	
	P100	2016	NVLink1	160	NVLink1	160	
AMD	MI300X	2023	Infinity Fabric Links（16x PCIe Gen5）	1024	16x PCIe Gen5	128	
	MI300A	2023		1024			
	MI250X	2021	Infinity Fabric Links（16x PCIe Gen4）	800	16x PCIe Gen4	64	
	MI250	2021		800			
	MI210	2022		300			
英特尔	Gaudi3	2024	16x PCIe Gen5	1200	16x PCIe Gen5	128	
	Gaudi2	2023	16x PCIe Gen4	600	16x PCIe Gen4	64	
Groq	LPU	2020	16x PCIe Gen5	600	16x PCIe Gen4	64	
大厂自研							
谷歌	TPU v6e	2024	ICI Links	800	/	/	/
	TPU v5p	2023	ICI Links	1200			
	TPU v5e	2023	ICI Links	400			
	TPU v4	2021	ICI Links	672			
	TPU v4i	2020	ICI Links	200			
亚马逊	Trainium2	2023	NeuronLink v3（PCIe Gen5）	1280	PCIe Gen5		
Meta	MTIA v2	2024	8x PCIe Gen5	64	8x PCIe Gen5	64	
	MTIA v1	2023	8x PCIe Gen4	32	8x PCIe Gen4	32	
微软	Maia 100	2023	8x PCIe Gen5	64	8x PCIe Gen5	64	

数据来源：各公司官网，A5 图王，量子位，新智元，AI 时代前沿，IT 之家，SemiAnalysis，半导体行业观察，智东西，芯智讯，机器之心，芯榜，科闻社，快科技，半导体产业纵横，电厂，电子工程专辑，东吴证券研究所

2. 为什么大厂纷纷开始自研 AI 芯片？——从自研成本测算说起

通常来说一个芯片公司的支出有以下四个方面：员工薪资、EDA 和 IP 费用、芯片制造费用、销售费用。以谷歌 TPU 与博通外包服务模式为例，这其中有部分由博通承担，但最终谷歌都需要支付相应的价格，因此我们不做口径调整，依然按 Fabless 公司的研发投入模式来计算。据老石谈芯对哲库造芯团队的研发投入测算，对于一家数字芯片 Fabless 公司而言，员工薪资约占总支出 60%，占掉大部分的比重。

员工薪资方面：1）研发人数：虽然互联网大厂自研与第三方芯片公司有一定的模式差异，但穿透下来 AI 芯片研发团队所需要的建制以及全流程粗算下来是可比的。从产品线来看，海外大厂英伟达的产品线丰富，且贯穿 AI 计算、AI 网络，也覆盖数据中心、游戏、汽车等诸多领域，与单纯做 AI 芯片的公司体量不可直接比较；我们认为，国内 AI 芯片&其他数字芯片公司的产品条线与之相对可比，我们以国内相关可比公司的研发人数来衡量大厂自研 AI 芯片所需研发人数（包括自身员工+外包服务商员工）。**2）人均薪酬：**2023 年度，寒武纪、海光信息、翱捷科技三家公司的研发人员平均薪酬为 82.13 万元，以此作为国内主流数字芯片设计研发人员的一般薪资水准。**由此测算，每年每团队所需员工薪资=研发人数×人均薪酬=1176 人×82.13 万元≈9.7 亿元。**这里只是薪资开支，其他还有福利等企业支出，所以总数会高于这个数字。

若一代产品的研发周期按 2 年计算，一代产品的研发投入可粗略计算=9.7 亿元*2/60%≈32.3 亿元，单卡价格以售价 1-1.5 万美元（A100 售价）即人民币 7-10.5 万元、毛利率 68.21%（英伟达 FY2023-FY2025 销售毛利率平均值）计算，**大约需要 4.5-7 万卡出货量可以覆盖前期的投入。**

图4：主流 AI 芯片公司研发人员数量情况

	公司名称	成立时间	研发人员			2023年研发人员人均薪酬（万元）
			2021	2022	2023	
688256.SH	寒武纪	2016-03-15	1213	1205	752	91.75
688041.SH	海光信息	2014-10-24	1031	1283	1641	85.44
688220.SH	翱捷科技	2015-04-30	914	991	1135	69.21
	平均值				1176	82.13

数据来源：各公司公告，iFinD，新浪财经，东吴证券研究所

头部大厂的万卡集群建设未曾停歇，完全有望覆盖自研 ASIC 的前期投入。1）训练端：从训练集群的规模上看，单一集群的需求量已逐渐超过 10 万卡。2023-24H1，各厂商陆续建成万卡集群，其中比较有代表性的是 Meta 于 24/03 月宣布的两个 24k GPU 集群（共 49152 个 H100）。24H2 以来市场最为关注的是 xAI 建设的 10 万卡 H100 集群，明年目标或将扩展至 100 万卡。**2）推理端：**英伟达 FY2024 数据中心有 40%的收入来自推理业务。随着 AI 应用遍地开花，我们认为 AI 推理需求还有更大渗透空间。

图5：主流科技公司公开宣布的万卡集群情况

厂商	算力建设			
	时间	GPU类型	GPU数量（万张）	备注
Microsoft	2020年	-	1.0	
Google	2023-05	H100	2.6	
	-	TPU v5p	0.9	
Meta	2022年	A100	1.6	
	2024年初	H100	2.5	共2个集群
	2024年底目标	H100	35.0	
AWS	2023-07	H100	2.0	
	正在部署	Trainium2	40.0	
xAI	2024年	H100	10.0	
	计划	-	100.0	

数据来源：消费日报网，机械之心，通信产业网，半导体行业观察，格隆汇 APP，东吴证券研究所

注：本图为非完全统计

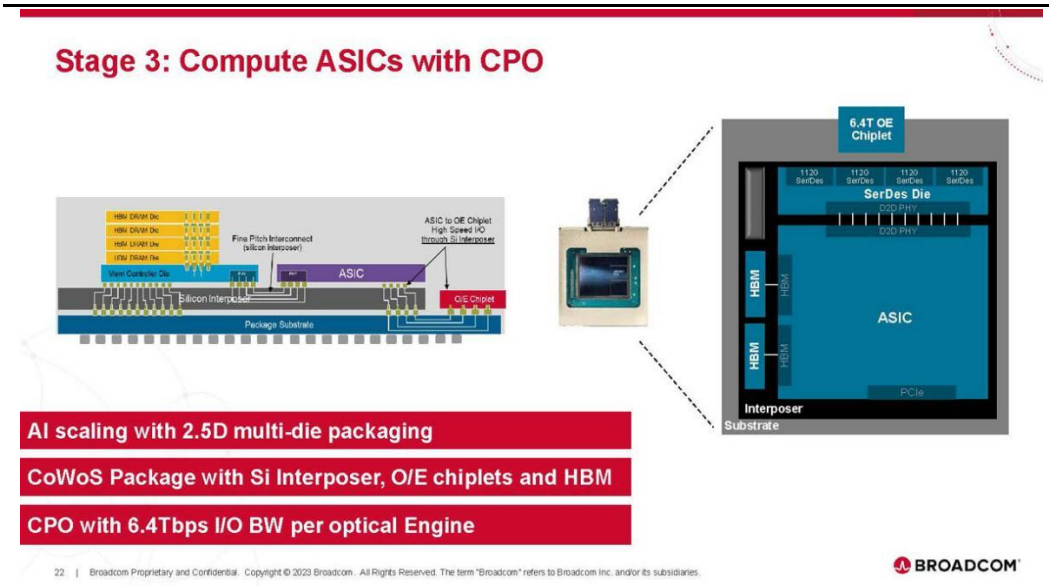
3. 大厂自研 AI 芯片谁能代工？

3.1. 博通：AI 互连技术引领者与半导体生态巨头

博通成立于 1991 年，是全球领先的 fabless 半导体设计与技术解决方案供应商，业务范围囊括多种半导体、企业用软件和安全解决方案的设计、开发和供应。

博通产品线 IP 生态强大完善，在接口、互连等领域保持前瞻性优势。接口方面，博通针对不同规模的 AI 集群提供差异化系统架构与解决方案。一类是 Endpoint Scheduled 系统架构，其主要面向的是小规模 AI 集群数据调度，各计算节点（如 GPU）之间通过 Tomahawk 5 以太网交换芯片来进行互联。从 2010 年的 640Gbps 增长到 2022 年的 51.2Tbps，Tomahawk 实现了 80 倍带宽提升，并且实现了超过 90%能耗降低。另一类是对于大规模 AI 集群、需要由智能交换机负责数据调度的 Switch Scheduled 架构中，博通使用上层 Spine 交换机 Ramon 和下层 Leaf 交换机 Jericho3-AI 来实现多路径互联，Jericho3-AI 芯片可连接多达 32,000 个 GPU，每个 AI 加速器能够提供 800Gbps 的数据带宽，最终能使网络性能提升 10%；在互连领域，2024 年博通发布了第二代网卡芯片 Thor 2，Thor 2 是业界首款采用 5nm CMOS 工艺实现的 400 千兆以太网（GbE）NIC 设备，支持 16 条 PCI Express 5.0 通道，每条通道的运行速度为 32 Gbps。而且 Thor 2 还可以直接驱动长达 5 米的铜缆，而大多数 NIC 竞争对手只能驱动 2.5 米长的铜缆。Thor 系列还通过 RoCE v2 在以太网上实现类 InfiniBand 性能，降低客户架构迁移成本。

图6：博通集成光学连接技术的 ASIC 芯片



数据来源：hotchips，东吴证券研究所

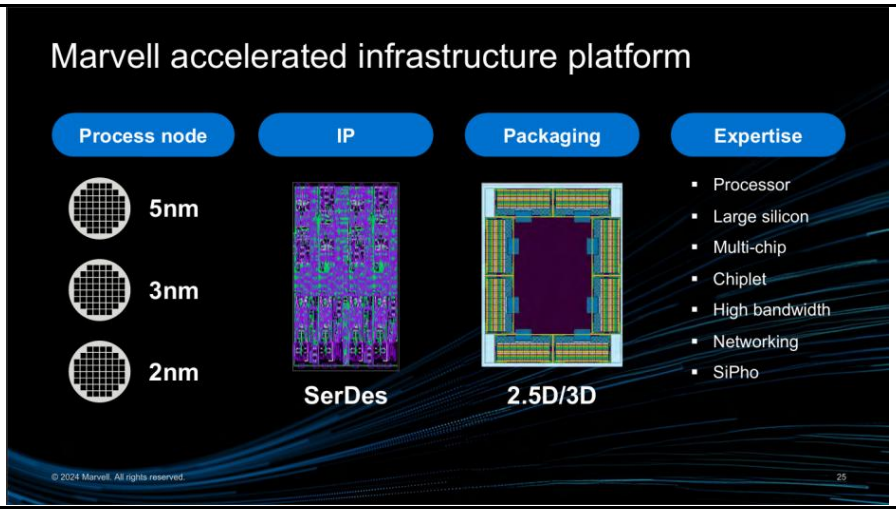
市场表现方面，博通营收动能强劲，AI 相关业务收入增长迅猛，合作范围稳步扩张。博通是全球最大的 AI 定制芯片服务商，根据博通今年的财报会所述，其在 175 亿美元的 TAM（可服务市场）中占到大约 70%的份额，这一部分也是博通快速增长的业务，而且市场潜力较大。2024 财年，博通 AI 相关收入达 122 亿美元，同比激增 220%，占半导体业务的 40%。2023 年第四季度网络业务收入 45 亿美元（同比增长 45%），其中 AI 网络收入占比 76%（同比增长 158%），原因包括对谷歌、Meta、亚马逊三家超大规模客户 AI XPU 出货量翻倍，同时全球范围内的 Tomahawk 与 Jericho 芯片出货量推动 AI 连接收入增长 4 倍；除此之外，博通宣布已被另外两家正在进行下一代自研 AI XPU 高级开发的超大规模供应商选中，2027 年博通的 AI ASIC 市场的机会将在 600 亿至 900 亿美元之间。

3.2. Marvell：数据中心芯片定制化赛道破局者

Marvell（美满电子）成立于 1995 年，早期以存储控制器技术立足半导体行业，2016 年全面转向数据中心芯片解决方案，聚焦 ASIC、光电器件、以太网交换芯片等领域。2024 年，公司营收达 39.5 亿美元，其中数据中心业务贡献 70%（27.98 亿美元），AI 相关收入占比从 2023 年的 5%跃升至 30%以上，成为其增长核心引擎。

Marvell 通过 HBM 重构与 CPO 集成的双重突破体现竞争力，直击 AI 芯片的能效与带宽瓶颈。通过非行业标准的 HBM I/O 接口设计，实现接口功耗降低 70%，将 HBM 支持电路从 XPU（AI 加速器）边缘移至堆叠底部的基础裸片，释放 XPU 芯片 25%的面积用于计算单元扩展。优化后，单一 XPU 可连接的 HBM 堆栈数量提升至高 33%，XPU 的性能和能效整体提高，降低了云运营商的 TCO。2025 年 1 月，Marvell 发布全球首款集成 CPO 的定制 XPU 架构，其 3D SiPho（硅光）引擎支持 200Gbps 电气/光学接口，集成 32 通道及光模块驱动单元，使单器件带宽与 I/O 密度翻倍，相比传统 100G 方案每比特功耗降低 30%。消除了电信号离开 XPU 封装进入铜电缆或穿过印刷电路板的需要，实现了 XPU 数据间传输的快速化与长距离化，相较电缆长 100 倍。

图7：Marvell 芯片制造各流程基础设施建设



数据来源：Marvell，东吴证券研究所

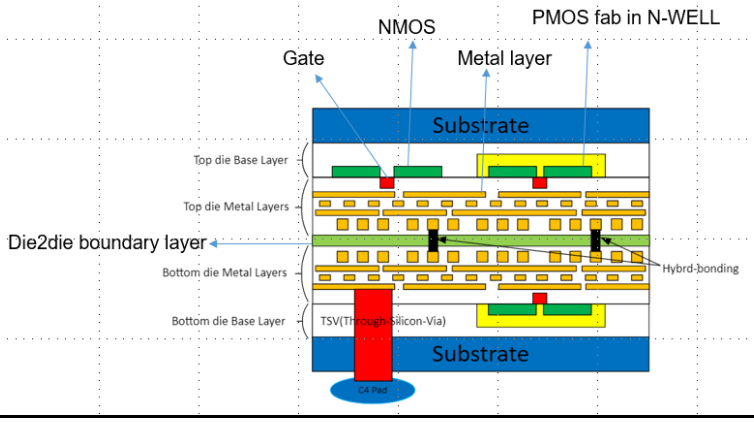
Marvell 对自研 AI 芯片的大厂市场深度渗透，在 AI 芯片市场扩张迅猛。Marvell 已与全球云端服务供应商（CSP）亚马逊 AWS 签署五年合作协议，向亚马逊 AWS 提供包括定制 AI 产品在内的系列芯片。此外，谷歌计划逐步放弃博通定制的 TPU，最早在 2027 年转向 Marvell，由它来定制新的 AI 芯片。CEO 还宣布 Marvell 已经获得了第三个 AI 超大规模客户并正在合作设计一款 AI 加速器，预计将于 2026 年投入生产，该客户所贡献的收入可能比与前两位客户的总和还要大。2024 年 Marvell 营业收入合计达 39.5 亿美元，其中数据中心业务营业收入达 27.98 亿美元，占总收入 70%，AI 业务增长显著，2023 年 12 月后，公司 AI 业务在总收入占比从 5%提升至 30%以上。根据公司及行业预期，Marvell 所在的 ASIC 行业 TAM 未来有望成长至 2028 年（2029 财年）的 429 亿美元。随着亚马逊相关芯片需求的增长以及其他两大客户的芯片产品出货，公司预期未来市占率有望达到两成。

3.3. AIchip：3DIC 与先进制程 ASIC 设计的先锋

世芯电子（AIchip）股份有限公司于 2003 年 2 月 27 日注册于英属开曼群岛，主要经营研究开发、设计及制造特殊应用积体电路设计(ASIC)和系统单晶片(SOC)及提供相关服务等。

世芯电子通过垂直堆叠与纳米片晶体管技术直击 AI 与高性能计算芯片的能效与算力瓶颈。世芯已正式开放其面向 AI 和高性能计算应用的最新高性能 ASIC 的三维集成电路（3DIC）设计服务，通过硅通孔（TSV）和混合键合技术实现多芯片垂直堆叠，世芯经过硅验证的 3DIC 设计流程从三个关键维度优化了选定的 3DIC 设计：功率传输、晶粒间电气互连和系统范围的热特性。与传统的二维设计相比，3DIC 可实现数据更快的传输、更低的功耗和更小的占用空间。同时，其 2nm 制程技术采用纳米片晶体管（GAA 架构），已完成测试芯片流片，集成并验证了该公司的 AP-Link-3D I/O IP，可用于 3DIC 小芯片系统设计。该流片结果为下一代 1.6nm 工艺技术的未来发展做好准备。二者的结合使世芯既能解决当下高算力场景的“功耗墙”难题，又能以异构集成能力定义下一代 AI 芯片架构，成为少数同时掌握先进制程与封装协同设计能力的全球领军者。

图8：AIchip 3DIC ASIC 芯片设计结构



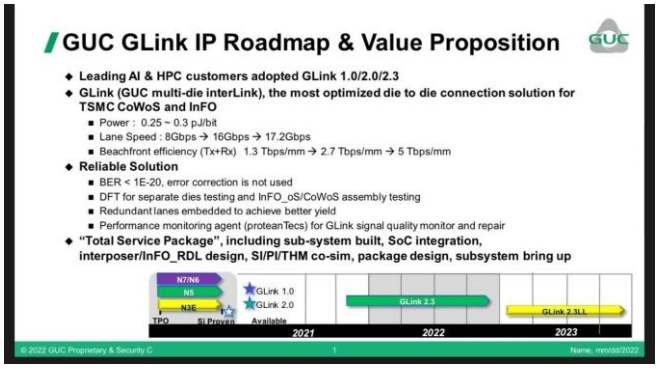
数据来源：世芯电子，东吴证券研究所

3.4. GUC：先进制程与封装覆盖的 ASIC 领导厂商

创意电子成立于 1998 年，专注于提供从 Spec-in 到 GDSII 交付的全流程 ASIC 设计服务，覆盖 0.5μm 至 3nm 等广泛工艺节点。

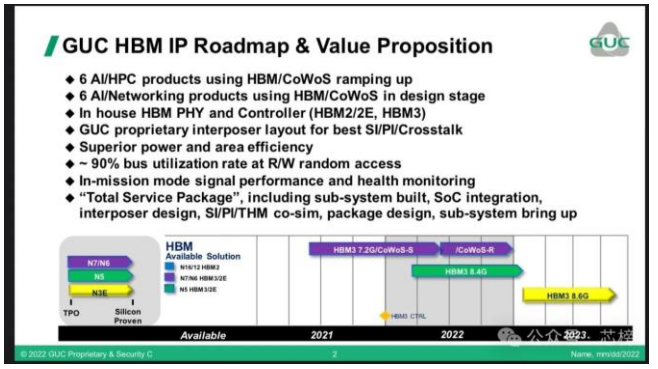
创意电子的技术竞争力体现在先进制程覆盖与系统级设计能力两大维度。创意电子提供的 SoC 设计服务涵盖了 0.5um 到先进的 5nm/4nm/3nm 等工艺节点制程，包括各种产品形态的芯片设计实现，完善的流程方法可以解决实现几十亿门级以上的规模设计、GHz 工作频率、深亚微米噪声耦合、IR 压降、静电(ESD)放电、可制造性设计(DFM)、良率设计提升(DFY)与严峻上市时程需求的挑战。创意电子的先进设计流程，增加了快速原型制作、自动化电源设计方案(考虑压降与可绕性的电源铺设，依照电流量与电压爬升率的电源开关并串接， 防范动态压降的电源密度分析)，有效减少时序收敛的设计流程迭代(针对不同温度/电压/制成的定制化方案 sign off，数据流分析工具， 时钟延迟 CDA 单元库定制，hold-free 工艺库定制等)，可制造性设计，及良率设计等已经通过数百个在深亚微米技术上第一次设计就成功的芯片设计客户项目的考验。此外，创意电子在可测试性设计（DFT）方面表现卓越，提供扫描电路合成、边界扫描、智能分组内存 BIST 测试等全套解决方案，确保芯片良率与可靠性。在异构集成领域，创意电子于 2025 年 1 月推出 UCIe™物理层 IP，支持每通道 40Gbps 的高速互连，适用于 AI、HPC 及网络芯片。该 IP 基于台积电 N5 制程，不仅超越现有 UCIe 标准速度，还通过先进封装技术（如 2.5D/3D 集成）实现多芯片系统的高效互联，为复杂计算场景提供底层支撑。

图9：创意电子互连技术发展



数据来源：创意电子，芯榜，东吴证券研究所

图10：创意电子存储技术发展



数据来源：创意电子，芯榜，东吴证券研究所

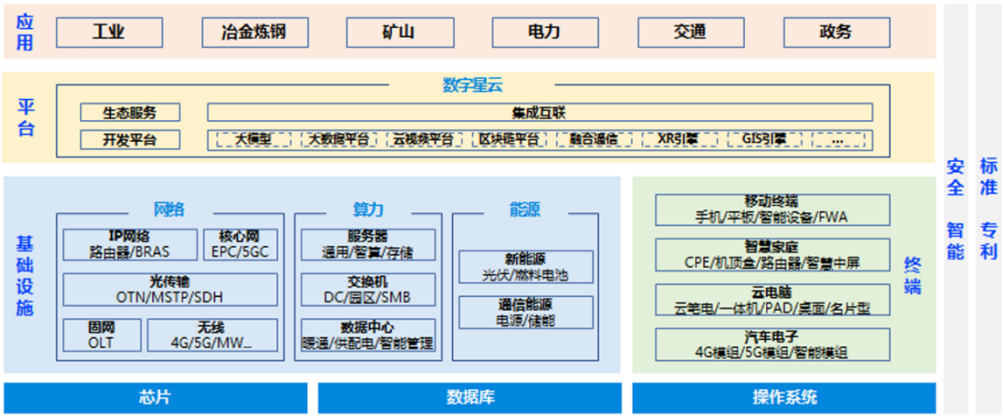
3.5. 中兴通讯：引领算力基础设施创新的国内大厂

中兴通讯作为全球知名的通信与信息技术解决方案提供商，在通信领域拥有深厚的技术积累和广泛的业务布局。其业务涵盖运营商网络、政企业务和消费者业务，在全球通信市场占据重要地位，为其在 ASIC 芯片相关业务发展提供了坚实基础。

在技术实力上，中兴通讯凭借多年研发沉淀，在算力基础设施领域掌握多项核心技术，构建起完备的技术体系。针对 AI、高性能计算等多样化场景需求及中心到边缘不同

层次的部署差异，中兴推出包括通算服务器、智算服务器、高性能存储以及训推一体机等的系列产品与方案全系列服务器支持液冷和异构加速，推出高密度全液冷整机柜解决方案，大幅降低数据中心能耗，提供分布式磁阵和全闪磁阵组合，兼顾存储大容量和高性能需求；根据 IDC 2023Q3 报告，中兴数据中心交换机国内市场份额同比增速第一，新一代高性能 400GE/800GE 数据中心交换机，支持单槽 14.4T，采用智能无损技术实现零丢包、低时延，具备绿色节能特性，助力客户打造架构优、功能全、高可靠的智算中心网络；创新框式单层多轨组网方案，可灵活高效构建千卡/万卡算力集群。该系列产品保持 GlobalData 国内同产品最高评级 Very Strong，其中 Hardware 单项获 Leader 最高评级。AI 相关业务贯穿了中兴通讯旗下运营商、政企、消费者三大业务线，中兴在其消费者业务中提出“AI for All”理念，布局全系 AI 终端。

图11：中兴通讯产品布局



数据来源：中兴通讯，东吴证券研究所

从市场表现来看，中兴通讯在国内外市场打下了广泛的市场根基。在国内运营商均实现规模项目交付；中标腾讯多层库框架项目，持续保持头部互联网的优势；建设中银宝信、中信银行等样板点，提升金融和数据中心行业市场的渗透率。在海外，聚焦重点国家，积极打造粮仓市场，独家承建印尼中国电信数据中心项目，突破其本土 DCI 数据中心运营商从而实现规模化经营，同时开拓埃塞、阿尔及利亚、利比亚等国家市场。

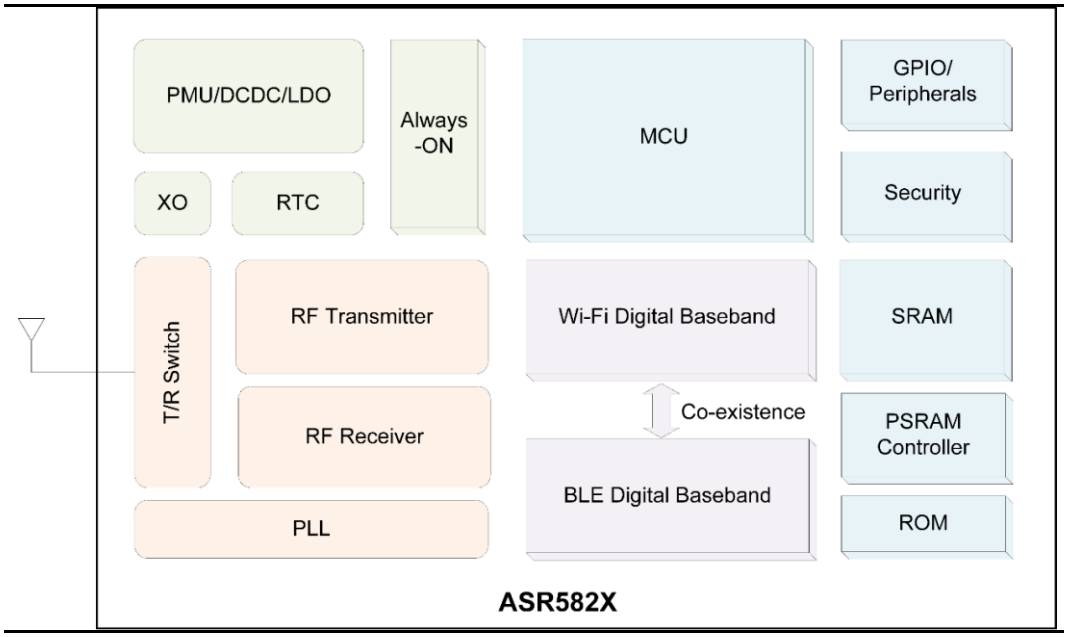
3.6. 翱捷科技：多方优势的平台型芯片企业

翱捷科技（ASR）是一家提供无线通信、超大规模芯片的平台型芯片企业。公司自设立以来一直专注于无线通信芯片的研发和技术创新，同时拥有全制式蜂窝基带芯片及多协议非蜂窝物联网芯片设计与供货能力，且具备提供超大规模高速 SoC 芯片定制及半导体 IP 授权服务能力。

翱捷科技具备完备齐全的自研 IP，在蜂窝基带芯片、非蜂窝物联网芯片设计与制造方面经验丰富。公司自主研发并积累了包含 2G 至 5G 的多模通信协议栈 IP、ISP、display、LPDDR2/3/4x、USB 2/3 Phy、PCIe Phy 等 SoC 芯片所需的大部分模拟 IP 及

数字 IP，可运用于各类芯片设计。ASR 已构建完整的多模全制式蜂窝基带产品布局，涵盖 Cat.1、Cat.4、Cat.6 和 5G 等全系列产品，品类齐全、竞争力强。ASR 的蜂窝基带产品广泛适用于物联网、智能手机和智能模组市场，能够满足各种通信终端的需求。目前，4G 蜂窝基带产品已在全球范围内实现持续量产出货，5G 蜂窝基带产品也在稳步推新中。ASR 拥有基于 Wi-Fi、LoRa、蓝牙技术的多种高性能非蜂窝物联网芯片，可广泛覆盖智能物联网市场各类传输距离的应用场景。

图12：翱捷科技 ASR582X 系列芯片图解



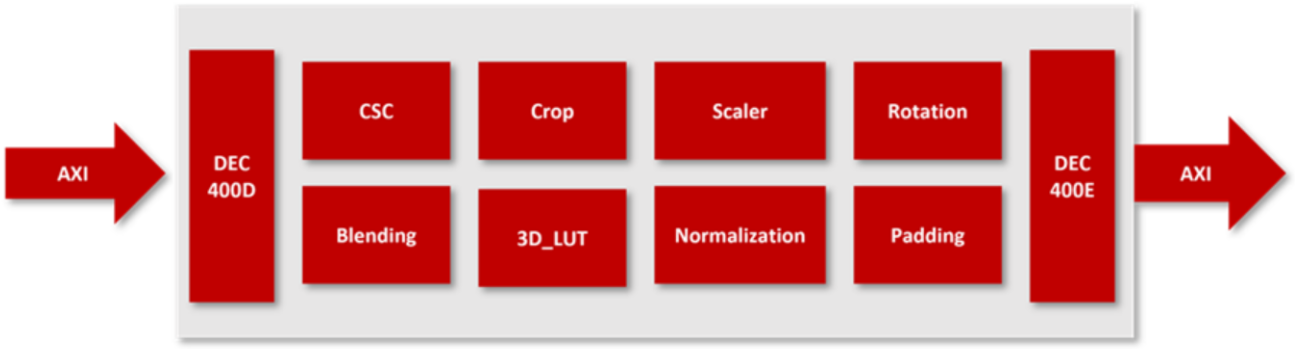
数据来源：翱捷科技，品慧电子网，东吴证券研究所

3.7. 芯原股份：平台化一站式芯片定制

芯原微电子（芯原股份）是一家依托自主半导体 IP，**为客户提供平台化、全方位、一站式芯片定制服务和半导体 IP 授权服务的企业。**在芯原独有的芯片设计平台即服务(Silicon Platform as a Service, SiPaaS) 经营模式下，通过基于公司自主半导体 IP 搭建的技术平台，芯原可在短时间内打造出从定义到测试封装完成的半导体产品，为包含芯片设计公司、半导体垂直整合制造商 (IDM)、系统厂商、大型互联网公司和云服务提供商在内的各种客户提供高效经济的半导体产品替代解决方案，业务范围覆盖消费电子、汽车电子、计算机及周边、工业、数据处理、物联网等行业应用领域。

芯原拥有多种芯片定制解决方案，包括高清视频、高清音频及语音、车载娱乐系统处理器、视频监控、物联网连接、智慧可穿戴、高端应用处理器、视频转码加速、智能像素处理等；此外，芯原还拥有 6 类处理器 IP，分别为图形处理器 IP、神经网络处理器 IP、视频处理器 IP、数字信号处理器 IP、图像信号处理器 IP 和显示处理器 IP，以及 1,500 多个数模混合 IP 和射频 IP。

图13：芯原视频后处理 IP PC820



数据来源：芯原微电子，东吴证券研究所

4. 风险提示

大厂 CapEx 投入不及预期。无论是 GPGPU 还是大厂自研的 ASIC 芯片，都需要 CapEx 投入的支持。若 AI 卡的应用需求不及预期引发大厂 CapEx 投入放缓，或影响 ASIC 进展。

技术发展不及预期。目前 ASIC 芯片算力性能、显存带宽与 GPU 仍有较大差距，或限制市场中 ASIC 芯片的应用范围。

客户需求不及预期。ASIC 下游主要客户厂商较为集中，单一客户需求变动对公司营收影响较大，若下游客户后期投资减少或增长缓慢，将影响芯片研发与产出进度。