东吴证券 SOOCHOW SECURITIES

电子行业点评报告

HBM 带宽限令对算力芯片有何影响？——算力芯片看点系列

2025 年 06 月 12 日

增持（维持）

证券分析师 陈海进
执业证书：S0600525020001
chenhj@dwzq.com.cn
研究助理 李雅文
执业证书：S0600125020002
liyw@dwzq.com.cn

投资要点

- **美国 HBM 带宽限令持续升级，对中国算力芯片产业构成显著挑战。**自 2022 年 10 月起，美国逐步将 HBM 纳入出口管制范围，并在 2024 年至 2025 年间多次加码，将管制标准精细化、范围扩大至非美国厂商，以限制尖端 HBM 技术对华出口，覆盖 HBM3e、HBM3 等最新尖端规格。2025 年 4 月，美国进一步将英伟达特供中国的 H20 芯片纳入禁令，彻底封锁中国获取高性能 AI 芯片的渠道。此次限令不仅针对 HBM 本身，还对显存带宽、互联带宽及两者带宽总和进行限制，使得国内企业难以通过技术降规或供应链调整做出应对。未来，美国 HBM 限令可能因为国内厂商使用的高性能 AI 芯片变化而进一步升级，对国产及海外进口 AI 芯片加以制裁。

- **HBM 带宽限制直接影响 AI 芯片性能，国内厂商短期面临算力压力。**从海外主流 AI 芯片解决方案来看，算力快速增长驱动 HBM 代际升级，带来带宽跃升。AI 芯片所需 HBM 显存带宽主要受到两个变量影响：1）芯片所能达到的算力峰值，与所需带宽成正比；2）芯片数据复用、缓存优化情况，与所需带宽成反比。根据主流芯片带宽利用率，可以合理估测其他芯片理论显存带宽需求。HBM 的高带宽和低延迟特性对 AI 训练和推理至关重要，限令导致国内企业难以获取先进 HBM，供应链面临中断风险。尽管企业通过算法优化、分布式训练和国产替代方案缓解部分影响，但短期内仍面临算力不足、推理速度下降等问题。此外，云端 AI 服务提供商可能因 HBM 短缺而被迫降低并发处理能力，影响用户体验。

- **长期来看，国产化替代和技术创新成为关键突破方向。**硬件方面，国内 AI 芯片厂商采取如 Chiplet 封装技术、GDDR6/LPDDR5 替代方案等应对方式，以降低对单芯片海外先进 HBM 的依赖。另外，还可对 HBM 数量进行减少、降规，或减少互联接口数量以降低显存带宽及互联带宽，从而满足新限令对于芯片带宽性能不得达到 H20 的要求。软件方面，国内云厂商积极推动软件优化，如低比特量化、模型剪枝，或转向低参数模型+分布式训练模式。此外，海外厂商可能推出降规版芯片（如英伟达特供 Blackwell）。但中国仍需加速自主 HBM 技术研发和产业链布局，以应对未来更严格的管制环境。

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

- **风险提示：**HBM 限令进一步升级风险，国产厂商研发进展不及预期的风险，关税政策风险

行业走势

相关研究

《电子行业中期策略——自主可控加码，AI 硬件加速落地》

2025-06-08

《海外 ASIC、GPU 双旺，国内传统行业 AI 转型迎来算力消耗拐点-算力周报》

2025-06-08

内容目录

图表目录

1. 美国 HBM 带宽限令变化梳理

1.1. HBM 出口限制起始时间与初始背景

美国 23 年首次将 HBM 纳入管制范围，随后不断细化升级层层加码。美国最早自 2022 年 10 月开始出台限制政策，旨在限制中国从外部获取及内部制造先进 AI 芯片的能力。2023 年 10 月，美国商务部工业与安全局（BIS）发布《先进计算和半导体制造物项规则》，要求出口商等在收到来自中国大陆/中国澳门等地区的订单时检视是否使用或包含 HBM，这是 HBM 作为 AI 芯片关键存储组件首次被纳入管制；2024 年 12 月，美国进一步修订管制条例，将技术参数从传统的技术节点（nm）细化升级，限制存储单元面积小于 0.0019μm²或存储密度大于 0.288Gb/mm²的 DRAM，从而控制本身就是多个 DRAM 通过 TSV 工艺堆叠封装的 HBM。2025 年 1 月，美国发布新一轮禁令，涵盖 HBM2e、HBM3、HBM3e 等尖端规格，基本封锁了中国获取最新 HBM 技术的渠道，同时加强供应链管控，限制使用美国技术的海外企业（如三星、SK 海力士）向中国供货。

图1：HBM 出口限令历史梳理

更改时间	具体内容	主要变动
2022.10	限制英伟达A100/H100等高端GPU对华出口，随后英伟达推出中国特供版A800/H800，主要通过降低互联速率来满足规定，仅小幅降低带宽。	首次直接限制中国获取外部先进AI芯片及内部制造能力，但并未直接限制HBM出口
2023.10	要求出口商等在收到来自中国大陆/中国澳门等地区的订单时检视是否使用或包含HBM	首次直接限制HBM
2024.12	限制存储单元面积小于0.0019μm²或存储密度大于0.288Gb/mm²的DRAM	HBM限制范围较以往标准更加精细化，直接限制DRAM面积及储存密度
2025.1	针对“memory bandwidth density”超过2GB/s/mm²的HBM产品，几乎覆盖当时所有量产型号，仅允许“与逻辑芯片合封”的HBM出口（如英伟达H20），同时限制SK海力士/三星的中国工厂生产先进HBM。	HBM限令扩大范围，涵盖非美国厂商生产的HBM2e、HBM3和HBM3e等尖端规格

数据来源：电子技术应用，中华网，韩民族日报，电子工程专辑，东吴证券研究所

1.2. 新一轮政策调整进一步扩大管制范围

限令不断升级趋势延续，H20 被美国政府新纳入管制范围。H20 作为英伟达此前在华唯一合规的专用 AI 芯片，填补了市场空白。然而，在 4 月实行的最新一轮管控中，对 HBM 和 I/O 带宽总和进行了限制，将原本特供中国的 H20 也纳入管制。美国时间 4 月 9 日，英伟达向美国证券交易委员会提交的 8-K 文件中详细披露了 H20 芯片禁令的

细节：禁止向中国（含港澳）及 D:5 国家（美国武器禁运名单）或总部位于这些国家/地区的公司，或其最终母公司所在国出口 H20 芯片，并且任何性能达到 H20 内存带宽或互连带宽，或达到两者组合性能的芯片均需许可证，以防止相关产品被用于或转用于中国的超级计算机。4 月 14 日，美国政府进一步通知英伟达，该许可要求将无限期有效，彻底永久限制中国获取 H20 芯片。

图2：英伟达公布内部文件截图

Item 8.01 Other Events.

On April 9, 2025, the U.S. government, or USG, informed NVIDIA Corporation, or the Company, that the USG requires a license for export to China (including Hong Kong and Macau) and D:5 countries, or to companies headquartered or with an ultimate parent therein, of the Company's H20 integrated circuits and any other circuits achieving the H20's memory bandwidth, interconnect bandwidth, or combination thereof. The USG indicated that the license requirement addresses the risk that the covered products may be used in, or diverted to, a supercomputer in China. On April 14, 2025, the USG informed the Company that the license requirement will be in effect for the indefinite future.

The Company's first quarter of fiscal year 2026 ends on April 27, 2025. First quarter results are expected to include up to approximately $5.5 billion of charges associated with H20 products for inventory, purchase commitments, and related reserves.

数据来源：英伟达，东吴证券研究所

1.3. H20 应用广泛&重要性高，新限令凸显美国 AI 战略

今年以来国产厂商积极采购，美国迅速禁用以限制中国 AI 进展。H20 是从英伟达 H200 裁剪而来，保留了 NVLink4.0 和 NVSwitch3.0 的 900GB/s 的卡间高速互联带宽，并支持 128GB/s 双向带宽的 PCIe Gen5。相较于英伟达 A800、H800 等，H20 在 FP8 算力、显存配置、卡间互联带宽、PCIe 连接等方面都有显著优势。2025 年前三个月，受 DeepSeek 低成本 AI 模型需求旺盛推动，H20 订单量激增，包括字节跳动和阿里巴巴在内的中国科技公司已大规模订购英伟达 H20 服务器芯片，总金额高达 160 亿美元，用于人工智能模型的训练与推理。由此可见，美国 HBM 限令聚焦中国云厂商的 AI 训练及推理，未来 HBM 限令可能因为国内厂商使用的高性能 AI 芯片变化而进一步升级，无论是国产还是海外进口 AI 芯片。

图3：H20 参数

	HGX H20	L20 PCIe	L2 PCIe
GPU Architecture	NVIDIA Hopper	NVIDIA Ada Lovelace	NVIDIA Ada Lovelace
GPU Memory	96 GB HBM3	48 GB GDDR6 w/ ECC	24 GB GDDR6 w/ ECC
GPU Memory Bandwidth	4.0 TB/s	864 GB/s	300 GB/s
INT8 | FP8 Tensor Core*	296 | 296 TFLOPS	239 | 239 TFLOPS	193 | 193 TFLOPS
BF16 | FP16 Tensor Core*	148 | 148 TFLOPS	119.5 | 119.5 TFLOPS	96.5 | 96.5 TFLOPS
TF32 Tensor Core*	74 TFLOPS	59.8 TFLOPS	48.3 TFLOPS
FP32	44 TFLOPS	59.8 TFLOPS	24.1 TFLOPS
FP64	1 TFLOPS	N/A	N/A
RT Core	N/A	Yes	Yes
MIG	Up to 7 MIG	N/A	N/A
L2 Cache	60 MB	96 MB	36 MB
Media Engine	7 NVDEC 7 NVJPEG	3 NVENC (+AV1) 3 NVDEC 4 NVJPEG	2 NVENC (+AV1) 4 NVDEC 4 NVJPEG
Power	400 W	275W	TBD
Form Factor	8-way HGX	2-slot FHFL	1-slot LP
Interconnect	Pce Gen5 x16: 128 GB/s NVLink: 900GB/s	PCIe Gen4 x16: 64 GB/s	PCIe Gen4 x16: 64 GB/s
Availability	PS: Nov 2023 MP: Dec 2023	PS: Nov 2023 MP: Dec 2023	PS: Dec 2023 MP: Jan 2024

数据来源：Wccftech，东吴证券研究所

2. HBM 显存带宽对 AI 芯片参数性能有何影响？

2.1. 主流 AI 芯片 HBM 配置

海外旗舰 AI 芯片普遍配置最新 HBM3e，算力快速增长驱动 HBM 代际升级，带来带宽跃升。以英伟达旗舰产品 GB200 为例，其采用 HBM3e 显存，实现 16384 GB/s 的超高带宽和 384GB 容量，较 H100（HBM3 3430 GB/s）带宽提升近 4 倍，容量增加 3.8 倍，下一代 GB300 预计将延续这一技术路线。类似地，AMD MI350X 也搭载 HBM3e，其 6144 GB/s 带宽较前代 MI300X（HBM3，5427 GB/s）提升显著。带宽的跃升式发展在历代 HBM 产品对比中更为明显：从 HBM2（如 V100 的 900 GB/s）到 HBM2e（A100 的 2039 GB/s），再到 HBM3/HBM3e，每代技术迭代都带来 50%-100%的带宽提升，从 HBM2 的 1000 GB/s 左右大幅跃升至 HBM3e 的 5000 GB/s 以上；显存容量也从 HBM2 时代的 32-80GB 大幅提升至 HBM3e 时代的 288-384GB，更好地满足了 LLM 等内存密集型应用需求。AI 芯片算力增长推动 HBM 配置不断升级迭代，通过高多层堆叠等技术创新，在单位面积带宽密度上实现突破，从而带来整体 HBM 带宽的跃升。

图4：国际主流方案 HBM 配置情况

				算力（TFLOPS/TOPS）										显存		
厂商	名称	发布时间	量产时间	FP64	FP64矩阵	FP32	FP32矩阵	TF32矩阵	FP16/FP16矩阵/BF16	INT8	FP8	FP6	INT4	FP4	HBM/显存	显存带宽(GB/s)	显存容量(GB)
海外芯片																	
英伟达	GB300														HBM3e		288
英伟达	GB200	2024/3/19	25Q3	90	90	180		2,500	5,000	10,000	10,000	10,000		20,000	HBM3e	16384	384
英伟达	H200	2023/11/13	24Q2	34	67	67		495	989	1,979	1,979				HBM3e	4915	141
英伟达	H100	2022/3/22	2022/9/21	34	67	67		495	989	1,979	1,979				HBM3	3430	80
英伟达	H800	2023/3/22	2023	1	1	67		495	989	1,979	1,979				HBM3	3430	80
英伟达	A100	2020/5/14	2020/5/14	10	20	20		156	312	624					HBM2e	2039	80
英伟达	A800	2022/11/8	22Q3	10	20	20		156	312	624					HBM2e	2039	80
英伟达	V100	2017/5/11		8		16			125	62					HBM2	900	32
AMD	MI350X	2024/6/3													HBM3e		288
AMD	MI325X	2024/6/3													HBM3e	6144	288
AMD	MI300X	2023/12/6	23Q4	82	163	163	163	654	1,307	2,615	2,615				HBM3	5427	192
AMD	MI300A	2023/12/6	2023/12	61	123	123	123	490	981	1,961	1,961				HBM3	5427	128
AMD	MI250X	2021/11/8		48	96	48	96		383	383			383		HBM2e	3277	128
AMD	MI250	2021/11/8	2021	45	91	45	91		362	362			362		HBM2e	3277	128
AMD	MI210	2022/3/22	2022/4	23	45	23	45		181	181			181		HBM2e	1638	64
英特尔	Gaudi3	2024/4/9	24Q3			14	229	459	1,835		1,835				HBM2e	3789	128
英特尔	Gaudi2	2022/5/1	2023/7												HBM2e	2519	96
Groq	LPU	2020/9/23							188	750					无，使用SRAM	81920	0.22
海外大厂自研																	
谷歌	TPU v6p																
谷歌	TPU v6e	2024/5/15	2024Q4						926	1,852					HBM3或3e	1640	32
谷歌	TPU v5p	2023/12/7	2024						459	918					HBM3	2765	95
谷歌	TPU v5e	2023/8/30							197	394					HBM2	819	16
谷歌	TPU v4	21Q4	2021Q4						138	275					HBM2	1228	32
谷歌	TPU v4i	20Q1	2021Q2						69	138						300	8
亚马逊	Trainium3	2024/12/3	2025年底						1,334	2,598							
亚马逊	Trainium2	2023/11/28	2024						667	1,299					HBM3	2900	96
Meta	MTIA v2	2024/4/10	2026			2.76	2.76		177	708					LPDDR5	205	128
Meta	MTIA v1	2023/5/18	2025			0.8	1.6		51	102					LPDDR5	176	64
微软	Maia 100	2023/11/16	2024						1,600	1,600				3,200	HBM2e	1843	64
特斯拉	D1	2021/8/20	2023/7	不支持	不支持	22.6	22.6		362	400					HBM	800	32

数据来源：英伟达、AMD 官网，中关村在线，雷科技，芯智讯，量子位，东吴证券研究所

国际先进 HBM 持续迭代升级，主流 AI 芯片配置继续提升。随着大模型对算力和存储带宽的需求激增，下一代 HBM 技术正迎来关键升级，主要聚焦于带宽提升、容量扩展、能效优化三大方向。HBM4 标准于 2025 年 4 月 17 日正式发布，其带宽显著提升。HBM4 支持 2048 位接口，传输速度高达 8 Gb/s，总带宽可达 2 TB/s，可满足 10 万亿参数模型需求。代工方面，SK 海力士采用台积电 3nm 工艺（原计划 5nm），三星用 4nm 工艺。三星已成功试产 16 层 HBM3 样品，计划用于 HBM4 量产。

图5：HBM 迭代产品参数

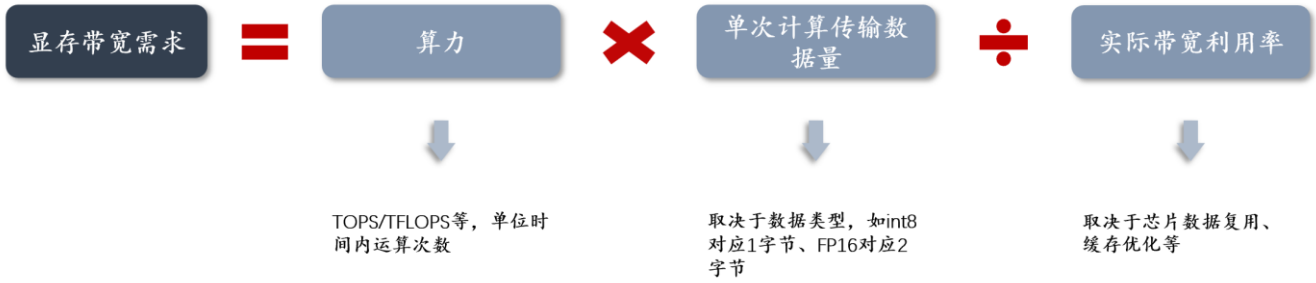
产品名称	时间	芯片密度（Gb）	带宽	堆叠高度（层）	容量（GB）	I/O速率（Gbps）	内存接口
HBM1	2014	2	128 GB/s	4	1	1	1024位
HBM2	2018	8	307 GB/s	4/8	4/8/16	2.4	1024位
HBM2e	2020	8/16	460 GB/s	4/8	8/16	3.6/3.2	1024位
HBM3	2022	16	819 GB/s	8/12	16/24	6.4	1024位
HBM3e	2024	24	1.2 TB/s	12/16	24/36	9.2	1024位
HBM4	2026	48	1.6/2 TB/s	16	36/64	/	2048位

数据来源：SK 海力士，美光，东吴证券研究所

2.2. HBM 显存带宽需求如何测算？

AI 芯片显存带宽需求可由算力及带宽利用率计算得出大致范围。从算力利用的角度看，AI 芯片所需要的 HBM 带宽可以简化为计算单元需要的数据吞吐量÷实际有效带宽利用率，需要确保在数据吞吐量达到峰值且实际带宽利用率处于较低水平时带宽仍能够喂饱计算单元，避免因 HBM 带宽不足而闲置算力。其中，数据吞吐量可以进一步分解为算力（TOPS）×每次计算传输数据量，数据传输量主要取决于数据类型，如 INT8=1 字节，FP16=2 字节，而实际带宽利用率代表计算单元实际需要从 HBM 读取的数据比例，主要受到数据复用率影响。因此，AI 芯片所需 HBM 显存带宽主要受到两个变量影响：1）芯片所能达到的算力峰值，与所需带宽成正比；2）芯片数据复用、缓存优化情况，与所需带宽成反比。

图6：HBM 显存带宽理论需求



数据来源：Haocheng Xi《COAT: Compressing Optimizer States and Activation for Memory-Efficient FP8 Training》，东吴证券研究所

根据主流芯片带宽利用率，可以合理估测其他芯片理论显存带宽需求。如下图所示，根据已知数据，可以计算得到各主流芯片理论最低带宽利用率。可以观察到随着迭代升级，带宽利用率呈提高趋势。因此，根据芯片本身存储优化情况可以大致估测带宽利用率，结合算力得到显存带宽需求。以英伟达 A100 芯片为例，其 FP16 算力达到 312 TFLOPS，假设其最低带宽利用率在 30-40%区间，因此对于显存带宽需求在 1560-2080 GB/s 区间，其使用的 HBM 带宽需要满足上述带宽要求。假设带宽需求取中值 1820 GB/s，对应 4 颗 HBM2e 或 3 颗 HBM3。

图7：理论最低带宽利用效率

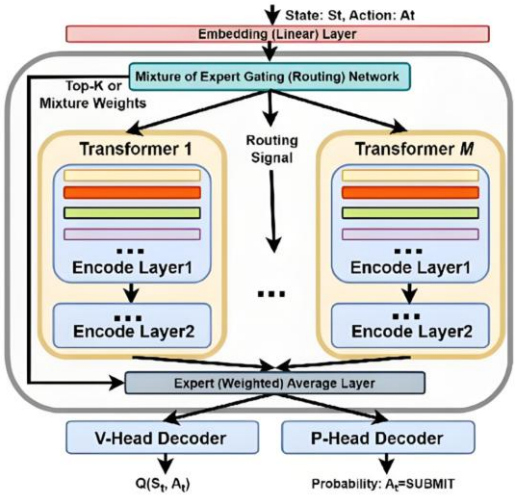
厂商	名称	发布时间	量产时间	FP16/FP16矩阵/BF16	INT8	HBM/显存	显存带宽(GB/s)	显存容量(GB)	理论最低带宽利用效率
海外芯片									
英伟达	GB300					HBM3e		288	
英伟达	GB200	2024/3/19	25Q3	5,000	10,000	HBM3e	16384	384	61%
英伟达	H200	2023/11/13	24Q2	989	1,979	HBM3e	4915	141	40%
英伟达	H100	2022/3/22	2022/9/21	989	1,979	HBM3	3430	80	58%
英伟达	H800	2023/3/22	2023	989	1,979	HBM3	3430	80	58%
英伟达	A100	2020/5/14	2020/5/14	312	624	HBM2e	2039	80	31%
英伟达	A800	2022/11/8	22Q3	312	624	HBM2e	2039	80	31%
英伟达	V100	2017/5/11		125	62	HBM2	900	32	28%
AMD	MI350X	2024/6/3				HBM3e		288	
AMD	MI325X	2024/6/3				HBM3e	6144	288	
AMD	MI300X	2023/12/6	23Q4	1,307	2,615	HBM3	5427	192	48%
AMD	MI300A	2023/12/6	2023/12	981	1,961	HBM3	5427	128	36%
AMD	MI250X	2021/11/8		383	383	HBM2e	3277	128	23%
AMD	MI250	2021/11/8	2021	362	362	HBM2e	3277	128	22%
AMD	MI210	2022/3/22	2022/4	181	181	HBM2e	1638	64	22%
英特尔	Gaudi3	2024/4/9	24Q3	1,835		HBM2e	3789	128	
英特尔	Gaudi2	2022/5/1	2023/7			HBM2e	2519	96	
Groq	LPU	2020/9/23		188	750	无，使用SRAM	81920	0.22	

数据来源：英伟达、AMD 官网，中关村在线，雷科技，芯智讯，量子位，东吴证券研究所

2.3. 限令对目前国内 AI 训练/推理具有一定冲击

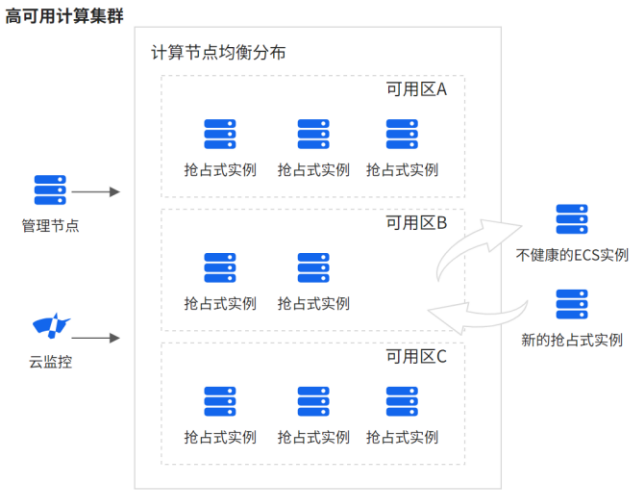
短期内企业采取训练算法优化、计算架构调整等方法，但限令仍带来较大算力压力。目前国内 AI 芯片厂商大多依赖三星、SK 海力士等出口的 HBM3 芯片，禁令后企业短期内无法获得足够 HBM，供应链面临中断风险。**算法方面，**阿里、字节等大厂主要通过优化算力分配在模型性能与算力需求间达成平衡。例如，使用 MoE 通过将任务分解给多个专门化的子模型或“专家”，由一个门控网络根据输入数据，动态选择合适的专家组合来处理特定任务，从而实现了计算效率与模型性能之间的平衡；阿里通过弹性伸缩实现自动批量创建 ECS 实例，同时使用均衡分布策略来实现将 ECS 实例均衡分散在多个可用区，并结合使用抢占式实例来降低算力资源成本。**计算架构方面，**企业通过异构计算及优化引入国产高性能算力卡集群，并针对性进行软件和架构优化，实现对现有英伟达算力卡的部分国产替代，但目前受限于性能及产能问题，未来 HBM 限令可能抑制国产算力卡产能提升，难以满足企业模型训练算力需求。

图8：混合专家模型（Mixture of Experts, MoE）



数据来源：阿里云开发社区，东吴证券研究所

图9：均衡分布策略



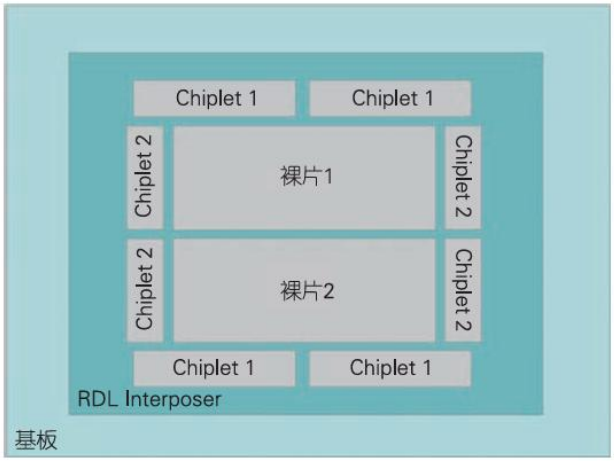
数据来源：阿里云帮助中心，东吴证券研究所

限令对 AI 推理芯片部署和应用具有一定影响，如推理速度下降、云端延迟增加。HBM 的低延迟特性能够显著减少数据传输时间，提高系统的整体响应速度，并且能够提供极高的数据传输速率，使芯片能够快速读取和写入大量数据，从而加速模型推理过程。HBM 禁令使得云端 AI 服务提供商难以获取足够的高性能 HBM 芯片，导致其不得不降低服务的并发处理能力，从而增加了云端服务的响应时间并降低了用户体验。

3. 限令影响 AI 训练/推理效果，国产厂商如何应对？

硬件方面，可使用 Chiplet 封装等技术实现技术跃升。由于美国 HBM 新一轮禁令主要限制显存带宽和互联带宽及其总和，要求均不得达到 H20 的参数，故国内厂商可以通过降低单芯片 HBM 数量、减配、减少互联接口数量等措施应对。具体而言，主要有三种可能的应对方式：1）减少 HBM3e 的数量，从而降低单芯片的显存带宽；2）减少 I/O 接口数量，从而降低接口总带宽；3）将 HBM3e 替换为较低版本的 HBM2，降低总显存带宽。另外，同样可以使用特殊封装技术以制造先进 AI 芯片，Chiplet 技术是一种利用先进封装方法将不同工艺/功能的芯片进行异质集成的技术。核心思想是先分后合，即先将单芯片中的功能块拆分出来，再通过先进封装模块将其集成为大的单芯片。不同芯粒可以根据需要来选择合适的工艺制程分开制造，再通过先进封装技术进行组装。

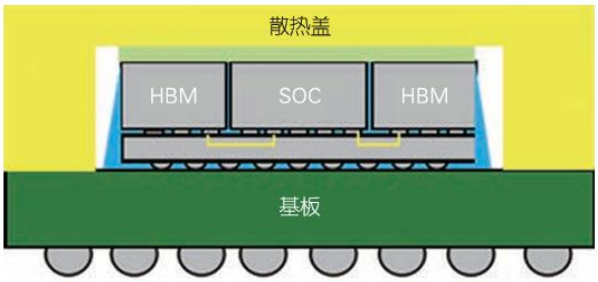
图10：2.5D（RDLInterposer）Chiplet 芯片封装正视图



RDL：重布线层

数据来源：李乐琪等《Chiplet 关键技术与挑战》，东吴证券研究所

图11：2.5D（RDLInterposer）Chiplet 芯片封装侧视图

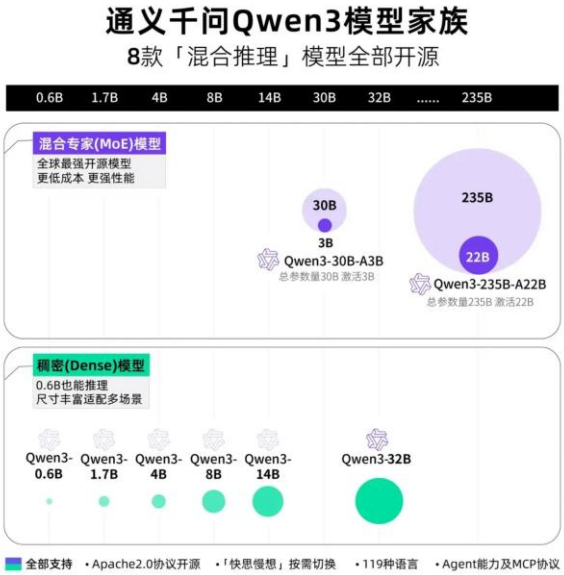


HBM：高带宽存储 SOC：片上系统

数据来源：李乐琪等《Chiplet 关键技术与挑战》，东吴证券研究所

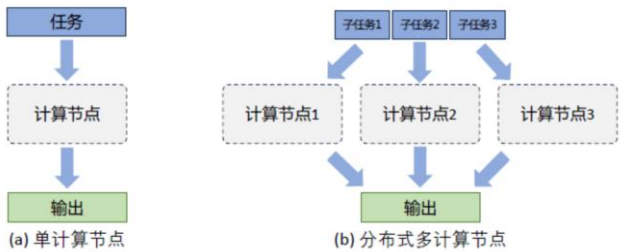
软件优化方面，国内云厂商可能选择转向低参数模型+分布式训练模式，以缓解算力瓶颈。低参数模型硬件对存储和计算资源的需求较低，更适合在现有硬件条件下运行，同时训练速度更快，能够在有限的硬件资源下更快地完成训练任务，对于需要快速迭代和优化的 AI 应用场景尤为重要，阿里最新发布天问 3.0 大模型参数量仅 2350 亿。另外，分布式训练将训练任务分解并分配到多个计算节点上共同完成，从而充分利用算力资源、提高训练速度和效率，从而能够训练更大、更复杂的模型，提高模型的精度和性能。

图12：阿里通义千问模型参数情况



数据来源：DOIT，东吴证券研究所

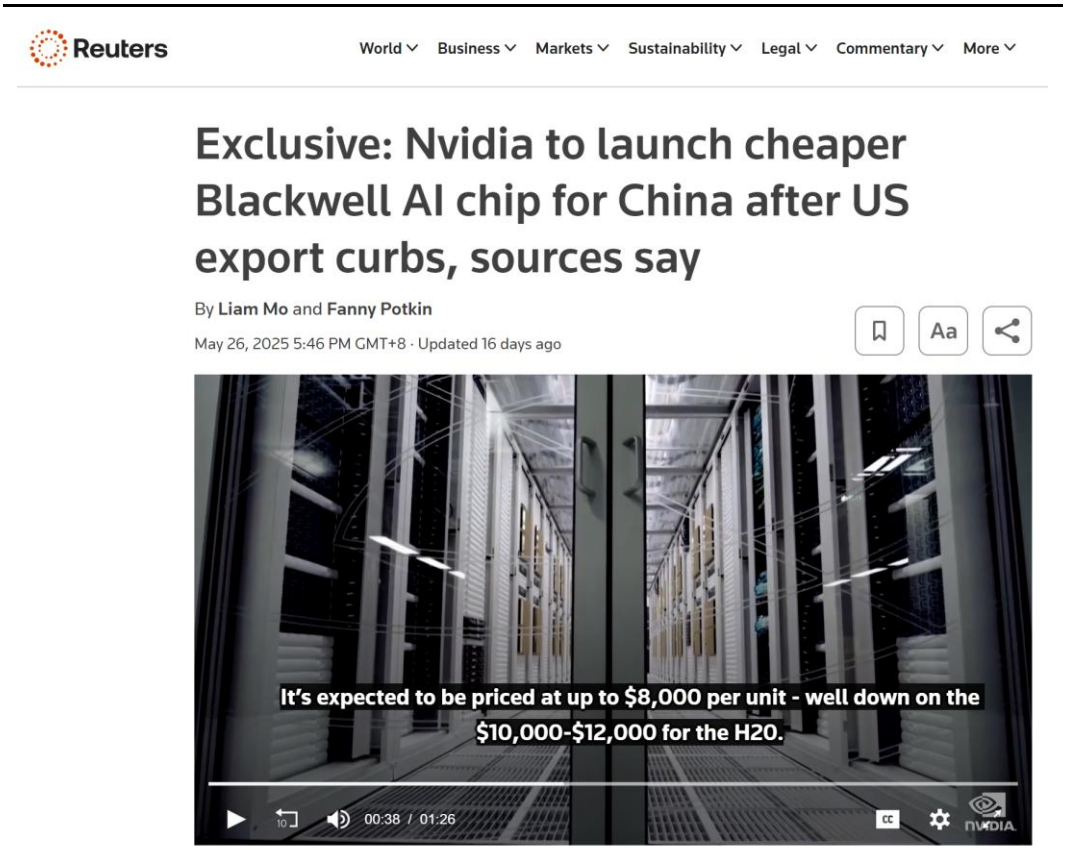
图13：分布式训练原理



数据来源：博客园，东吴证券研究所

另外，以英伟达为代表的海外厂商可能采取降规降价的方式应对。未来海外厂商可能推出更多降规版 AI 芯片以符合 HBM 限令要求，具体方式可能有使用非 HBM 内存，如 GDDR 等进行替代，或可用于 AI 推理，减轻国内云厂商算力不足的压力。**应对限令冲击，英伟达计划推出降规版 Blackwell 芯片特供中国。**据路透社消息，英伟达计划推出中国特供版 Blackwell 芯片，价格在 6500-8000 美金，低于 H20 的 10000-12000 美金。此款芯片将基于英伟达 RTX Pro 6000D，并使用 GDDR7 内存，以此配合美国对 HBM 的出口管制要求，另外新款芯片不会采用台积电 CoWoS 先进封装技术。

图14：英伟达降规版 Blackwell 芯片相关报道



数据来源：Reuters，东吴证券研究所

4. 风险提示

HBM 管制进一步升级的风险。美国 HBM 相关管制措施呈现持续升级态势，若限令进一步升级，或进一步限制国内 AI 芯片厂商获取高端 HBM。

国产厂商研发进展不及预期的风险。目前国内部分 AI 芯片采用国产 HBM 产品，若国产 HBM 厂商技术研发不及预期，将减少相关产品供给。

关税政策升级的风险。若美国加大关税征收力度，半导体产业链上游设备及材料进口或受到影响，提高 AI 芯片厂商制造成本。

东吴证券 SOOCHOW SECURITIES

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。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。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中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投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn