行业研究

行业专题报告

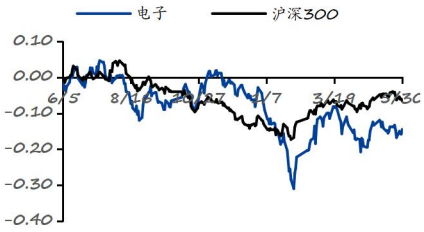
证券研究报告

电子

如何测算文本大模型 AI 训练端算力需求？

强于大市（维持评级）

一年内行业相对大盘走势



团队成员

分析师： 任志强(S0210524030001)
rzq30466@hfzq.com.cn
联系人： 徐巡(S0210124040079)

相关报告

1、半导体板块再度活跃，消费回暖趋势进一步明确-半导体系列跟踪——2024.06.03
2、【华福电子】20240531 周报：AI 终端崭露头角，Mini LED 或迎机遇——2024.06.01
3、巨头轮番入场，AI PC 爆发在即——消费电子系列跟踪——2024.06.01

投资要点：

➢ **需求侧：Scaling Law 驱动大模型算力需求不减**

Scaling Law 仍然是当下驱动行业发展的重要标准。Scaling Law 的基本原理是，模型的最终性能主要与计算量、模型参数量和数据大小三者相关，当不受其他两个因素制约时，模型性能与每个因素都呈现幂律关系。因此，为了提升模型性能，模型参数量和数据大小需要同步放大。从大模型数量上看，近年来呈现爆发式增长趋势。且由于尖端 AI 模型对于资源投入的大量需求，产业界对于大模型的影响力逐步加深。我们统计了产业界诸多公开披露的大模型训练数据，从大模型算力需求来看，GPT-3 到 GPT-4 参数上从 175B 快速提升到 1.8TB（提升 9 倍），训练数据量（Token 数）同方向快速增长，由 0.3TB 提升至 13TB（提升 42 倍）。绝对值上看，根据我们的非完全统计情况，国内外主流大模型在参数量上基本已来到千亿量级，在预训练数据规模上均已来到个位数乃至十位数的 TB 量级。

➢ **供给侧：黄氏定律推动英伟达 GPU 一路高歌**

英伟达 GPU 持续引领全球 AI 算力发展，虽然“摩尔定律”逐步放缓，但“黄氏定律”仍在支撑英伟达 GPU 算力快速提升，一方面，英伟达寻求制程工艺迭代、更大的 HBM 容量和带宽、双 die 设计等方法，另一方面，数据精度的降低起到关键作用，Blackwell 首度支持 FP4 新格式，虽然低精度可能会存在应用上的局限性，但不失为一种算力提升策略。若仅考虑英伟达 FP16 算力，A100/H100/GB200 产品的 FP16 算力分别为前代产品的 2.5/6.3/2.5 倍，在数量级上持续爆发，自 2017 年至今，GB200 的 FP16 算力已达到 V100 的 40 倍。与之对比，AI 大模型参数的爆发速度相对更快，以 GPT 为例，2018 年至 2023 年，GPT 系列模型从 1 亿参数规模大幅提升至 18000 亿。相较于 AI 大模型由 Scaling Law 驱动的参数爆发，GPU 算力增速仍亟待提升。

➢ **结论：预计 24-26 年全球文本大模型训练卡需求为 271/592/1244 万张**

我们根据侧算力供给需求公式，需求侧假设行业依然沿 Scaling Law 发展方向进一步增长，供给侧通过对英伟达 GPU 的 FP16 算力、训练市场、算力利用率等进行假设，推导得出 GPU 需求量。以英伟达 Hopper/Blackwell/下一代 GPU 卡 FP16 算力衡量，我们认为 2024-2026 年全球文本大模型 AI 训练侧 GPU 需求量为 271/592/1244 万张。

➢ **建议关注**

- 算力芯片：寒武纪 海光信息 龙芯中科

- 服务器产业链：工业富联 沪电股份 深南电路 胜宏科技

➢ **风险提示**

AI 需求不及预期风险、Scaling Law 失效风险、GPU 技术升级不及预期的风险、测算模型假设存在偏差风险。

正文目录

图表目录

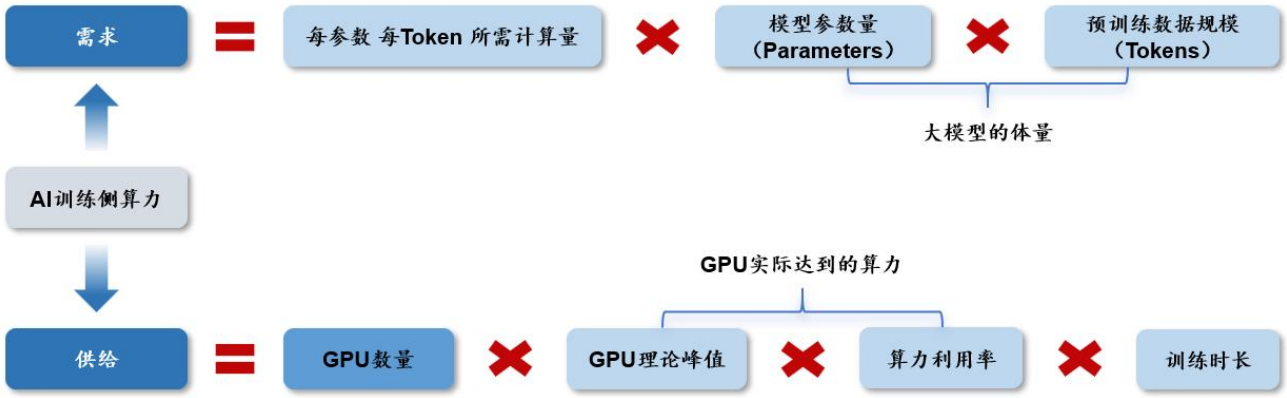
1 如何测算文本大模型 AI 训练侧算力需求？

对于 AI 训练侧算力，我们核心需要解决的问题是——当前蓬勃发展的 AI 大模型应用，到底带来多少 GPU 需求量。我们整理出算力供给需求公式，并分类讨论公式中的核心参数变化趋势，以此给出我们的判断。基于初步分析，我们将核心需要解决的问题进一步拆解如下：

1、**需求侧**，单个大模型训练计算量是否仍有提升空间？大模型数量如何演变？

2、**供给侧**，GPU 在实际应用中性能提升速度如何？

图表 1：文本大模型 AI 训练侧算力供给需求公式



来源：NVIDIA&Stanford University&Microsoft Research《Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM》，新智元，CIBA 新经济，华福证券研究所

从大模型算力需求来看，GPT-3 到 GPT-4 历时三年代际升级效果显著。参数上从 175B 参数快速提升到 1.8TB 参数（提升 9 倍），训练数据量（Token 数）同方向快速增长，由 0.3TB 提升至 13TB（提升 42 倍）。绝对值上看，根据我们的非完全统计情况，国内外主流大模型在参数量上基本已来到千亿量级，在预训练数据规模上均已来到个位数乃至十位数的 TB 量级。

从 GPU 供给端来看，算力利用率稳步提升，不同芯片种类之间体现出差异。GPT-3 到 GPT-4 明显看到算力利用率由 21.3%提升至 34%（32-36%区间，本文取中值粗略计算），趋势上较为明确。横向对比发现，相较于 OpenAI 的 GPT 系列，谷歌利用 TPU 训练的 Gropher 和 PaLM 明显在算力利用率上更胜一筹，我们认为谷歌自研 TPU 在自有大模型训练上展现出独特的优势。

图表 2：海外主流 AI 大模型训练侧算力供给需求情况

	单位	GPT	GPT2	GPT3	GPT4	Gopher	PaLM	PaLM 2
基本信息								
发布机构		OpenAI	OpenAI	OpenAI	OpenAI	DeepMind	Google	Google
发布时间		2018-06	2019	2020-05	2023-03	2021-12	2022-04	2023-05
AI训练								
大模型算力需求								
参数量	亿	1	15	1746	18000	2800	5400	3400
预训练数据规模（token）	万亿			0.3	13.0	0.3	0.8	3.6
GPU算力供给								
GPU产品				V100	A100	TPU v3	TPU v4	
GPU数量	片			10000	25000	4096	6144	
理论峰值FP16 TC	TFlops			125	312			
算力利用率				21%	34%	33%	46%	

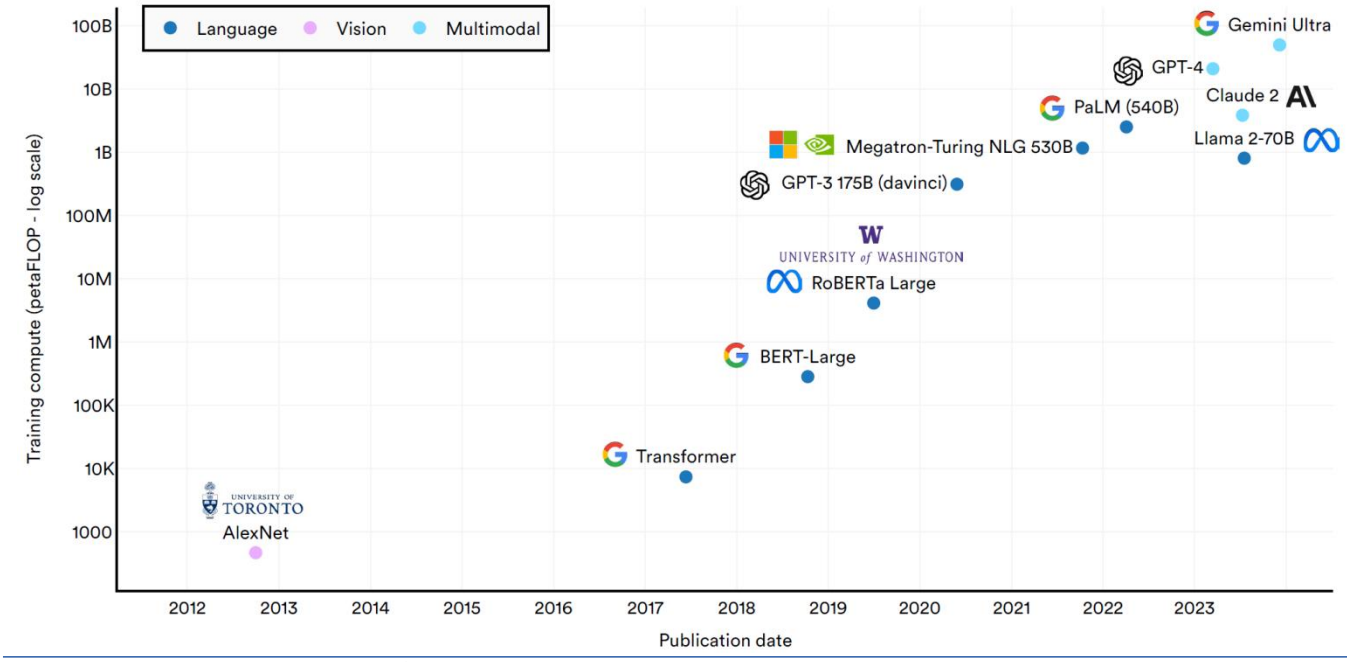
来源： OpenAI《Language Models are Few-Shot Learners》，Google《PaLM : Scaling Language Modeling with Pathways》，英伟达，谷歌研究院，腾讯科技，机器之心，中关村在线，河北省科学技术厅，华福证券研究所
注 1：由于各公司对于大模型的训练数据披露口径不一，以上为本文非完全统计
注 2：GPT4 算力利用率在 32-36%区间，本文取中值粗略计算
注 3：英伟达 V100 理论峰值为官网所示“深度学习 |NVLink 版本”性能

图表 3：国内主流 AI 大模型训练侧算力供给需求情况

	单位	混元	混元	Qwen-72B	Qwen1.5-110B	商量2.0	商量5.0	DeepSeek v2
基本信息								
发布机构		腾讯	腾讯	阿里	阿里	商汤	商汤	幻方
发布时间		2023-09	2024-04	2023-11	2024-04	2023-07	2024-04	2024-05
AI训练								
大模型算力需求								
参数量	亿	超千亿	万亿	720	1100	1040	6000	2360
预训练数据规模（token）	万亿	2.0		3.0		1.6	10.0	8.1

来源：腾讯云，通义千问公众号&GitHub 网页，新闻晨报，市界，IT 之家，华尔街见闻，新浪科技，钛媒体，华福证券研究所
注 1：由于各公司对于大模型的训练数据披露口径不一，以上为本文非完全统计
注 2：腾讯混元参数量披露口径较为模糊，分别为超千亿参数/万亿参数，在本图中不涉及左侧第二列单位

图表 4：各类别主流机器学习模型计算量



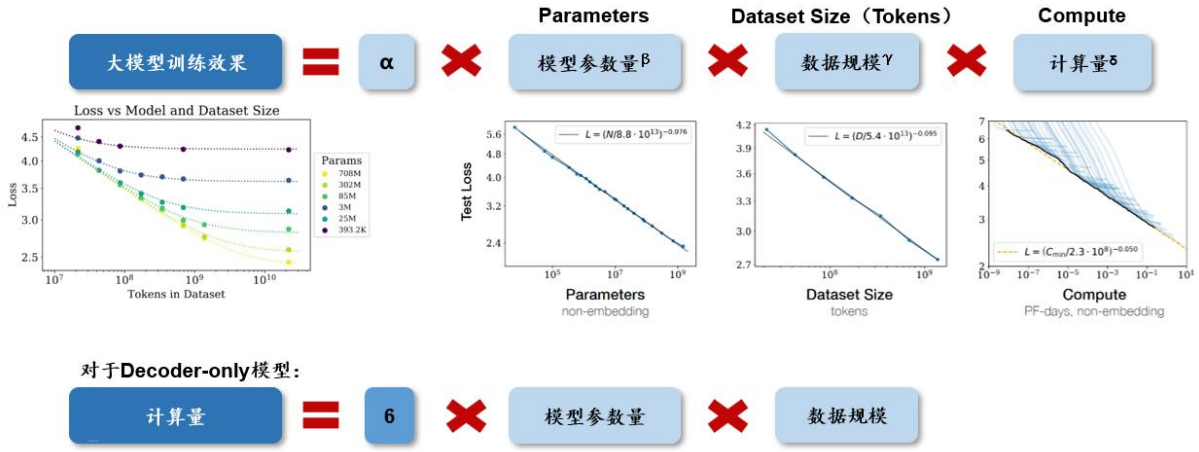
来源：HAI《2024 AI Index Report》，华福证券研究所

2 需求侧：Scaling Law 驱动大模型算力需求不减

2.1 Scaling Law 带动大模型参数爆发

Scaling Law 是深度学习领域的一个重要概念，它描述了模型性能与模型规模（如参数数量、数据规模和计算资源）之间的关系。Scaling Law 的基本原理是，模型的最终性能主要与计算量、模型参数量和数据大小三者相关，而与模型的具体结构（层数/深度/宽度）基本无关。如下图所示，对于计算量、模型参数量和数据规模，（1）当不受其他两个因素制约时，模型性能与每个因素都呈现幂律关系。（2）如模型的参数固定，无限堆数据并不能无限提升模型的性能，模型最终性能会慢慢趋向一个固定的值。**因此，为了提升模型性能，模型参数量和数据大小需要同步放大。**

图表 5：大模型训练的 Scaling Law



来源：OpenAI《Scaling Laws for Neural Language Models》，PaperWeekly，Expara Academy，华福证券研究所

Scaling Law 仍然是当下驱动行业发展的重要标准。假定计算量整体放大 10 倍，不同厂商在参数量和数据规模上各有权衡。OpenAI 认为模型参数更重要，模型参数应放大 $10^{0.73}$（5.32）倍，数据放大 $10^{0.27}$（1.86）倍；后来 DeepMind 和 Google 分别在 Chinchilla 和 PaLM 模型的工作中，认为模型参数量与数据同等重要，两者都应该分别放大 $10^{0.5}$（3.16）倍。其中，DeepMind 提出的 data scaling law（也称为 Chinchilla 或 Hoffman scaling laws）认为应该使用 1,400B（1.4T）tokens 来训练参数量大小为 70B 的大语言模型最佳。若在 Chinchilla scaling laws 的基础之上推断，单位参数大约需要 20 个 token 来进行训练。

图表 6：与 Chinchilla 数据优化模型一致所需的数据集大小

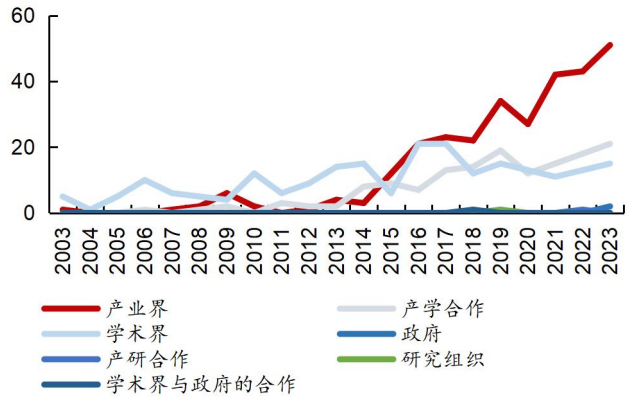
Model size (params)	Training tokens (round)	Training data used (estimate)	How much data is that? If 1 book is about 500KB of text (estimate)
Chinchilla/			*More books than in...*[23]
70B	1.4 Trillion	2.3TB	*The Kindle store on Amazon US (6.4M).*
250B	5 Trillion	8.3TB	*All 30 libraries at Yale University (16.6M).*
500B	10 Trillion	16.6TB	*The Google Books collection (33.2M).*
1T	20 Trillion	33.3TB	*The US Library of Congress (66.6M).*
10T	200 Trillion	333TB	*All US public libraries combined (666M).*
100T	2 Quadrillion	3.3PB	*All bibles ever sold worldwide (6.6B).*
250T	5 Quadrillion	8.3PB	*A stack all the way to the Moon (16.6B).*
500T	10 Quadrillion	16.6PB	*4 books about every living human (33.2B).*

来源：Alan D. Thompson《Chinchilla data-optimal scaling laws: In plain English》，华福证券研究所

2.2 大模型厂商持续涌现，AI 大模型数量激增

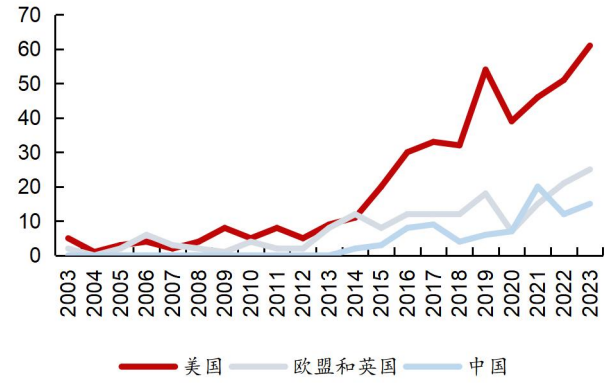
大模型数量呈现爆发式增长趋势，23 年产业界大模型突出重围。在 2014 年之前，学术界一直引领着机器学习模型的发布。此后产业界开始兴起，到目前已经逐渐来到领跑者的位置。2023 年，产业界发布了 51 个知名机器学习模型，而学术界仅发布了 15 个。现在，尖端 AI 模型需要大量的数据、计算能力和投资，而这些都是学术界所不具备的。从地区划分来看，根据研究人员所属机构所在地划分，2023 年，美国以 61 个知名机器学习模型居首，中国以 15 个紧随其后，欧洲市场总体模型数量之和略高于中国。

图表 7：各领域知名机器学习模型数量



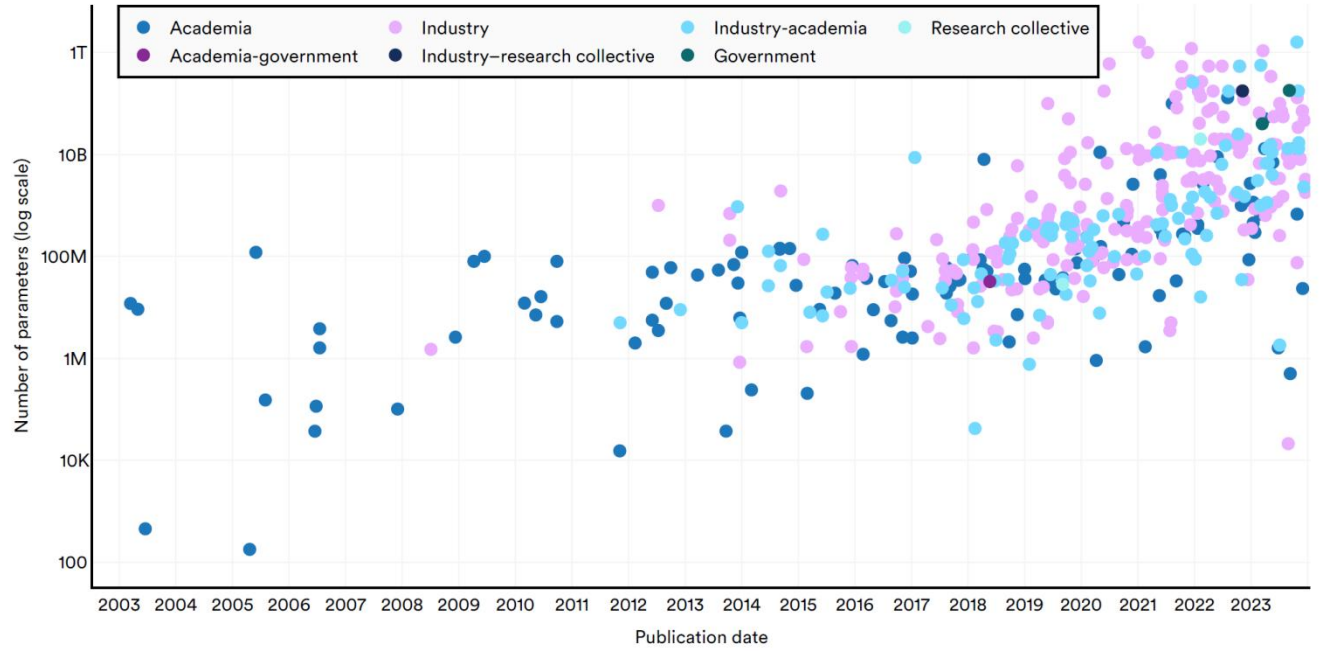
来源：HAI《2024 AI Index Report》，华福证券研究所

图表 8：各地区知名机器学习模型数量



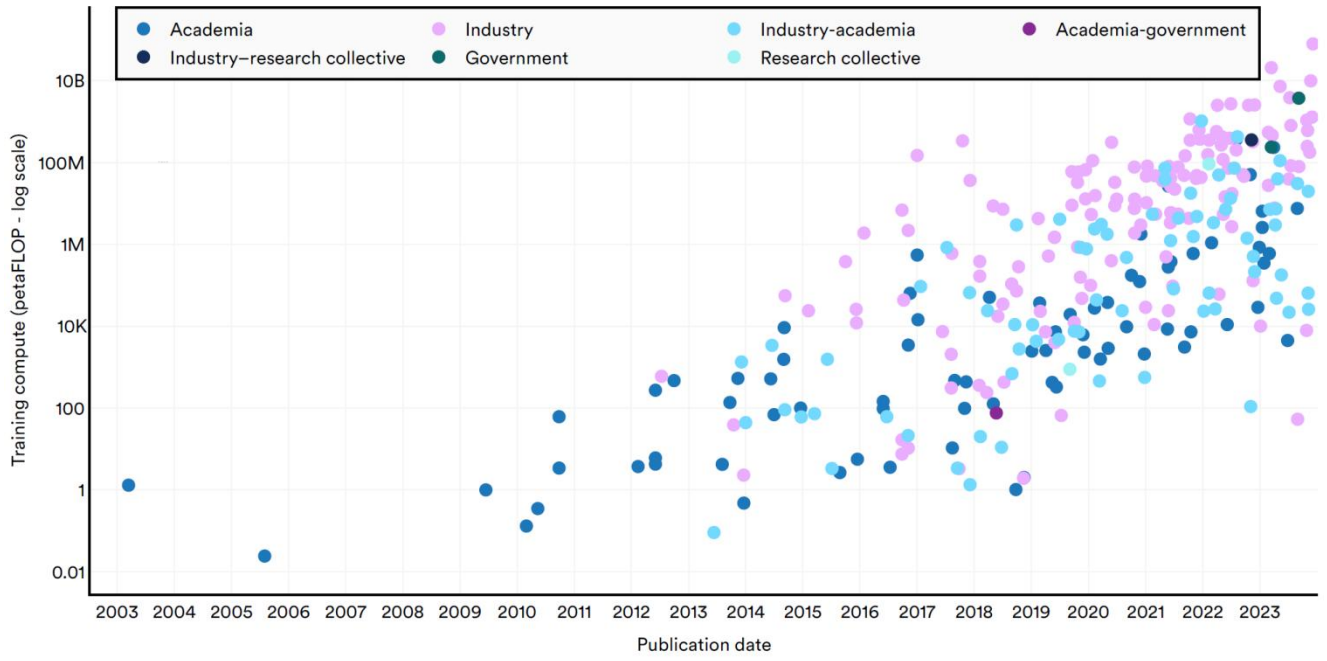
来源：HAI《2024 AI Index Report》，华福证券研究所
注：根据研究人员所属机构所在地划分

图表 9：各领域主流机器学习模型参数量



来源：HAI《2024 AI Index Report》，华福证券研究所

图表 10：各领域主流机器学习模型计算量



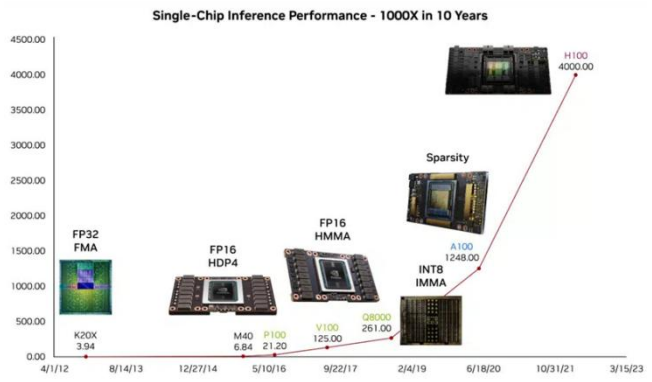
来源：HAI《2024 AI Index Report》，华福证券研究所

3 供给侧：黄氏定律推动英伟达 GPU 一路高歌

3.1 GPU：算力底层硬科技，支撑 AI 大模型发展

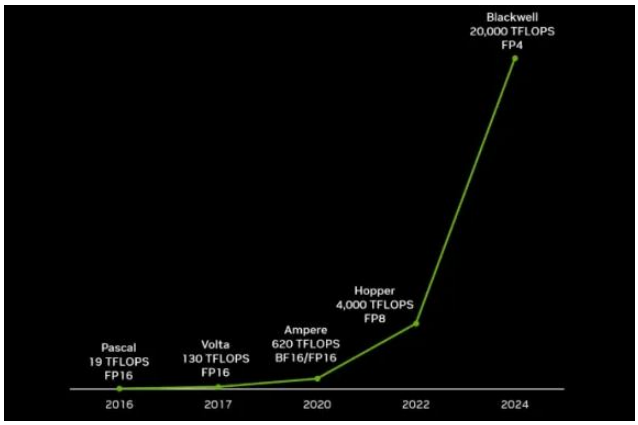
作为 GPU 行业龙头，英伟达“黄氏定律”持续奏效。英伟达 23Q3 发文，在过去十年中，英伟达 GPU AI 处理能力增长了 1000 倍。虽然“摩尔定律”逐步放缓，但“黄氏定律”意味着“单芯片推理性能”中看到的加速不会逐渐消失，而是会继续显现。24Q1 GTC 大会上英伟达进一步披露，从 16 年 Pascal GPU 的 19TFlops 到 24 年 Blackwell GPU 的 20PFlops，英伟达用 8 年将单卡 AI 训练性能提升了 1000 倍。

图表 11：英伟达 AI 性能提升-10 年 1000 倍



来源：英伟达，华福证券研究所

图表 12：英伟达 AI 性能提升-8 年 1000 倍



来源：英伟达，机器之心，华福证券研究所

除了得益于制程工艺迭代、更大的 HBM 容量和带宽、双 die 设计外，数据精度的降低起到关键作用，Blackwell 首度支持 FP4 新格式。多数训练是在 FP16 精度下进行，但实际上不需要用这么高的精度去处理所有参数。英伟达一直在探索怎么通过混合精度操作来在降低内存占用的同时确保吞吐量不受影响。Blackwell GPU 内置的第二代 Transformer 引擎，利用先进的动态范围管理算法和细粒度缩放技术（微型 tensor 缩放）来优化性能和精度，并首度支持 FP4 新格式，使得 FP4 Tensor 核性能、HBM 模型规模和带宽都实现翻倍。降精度的难点是兼顾用户对准确率的需求。FP4 并不在什么时候都有效，英伟达专门强调的是对混合专家模型（MoE）和大语言模型（LLM）带来的增益。

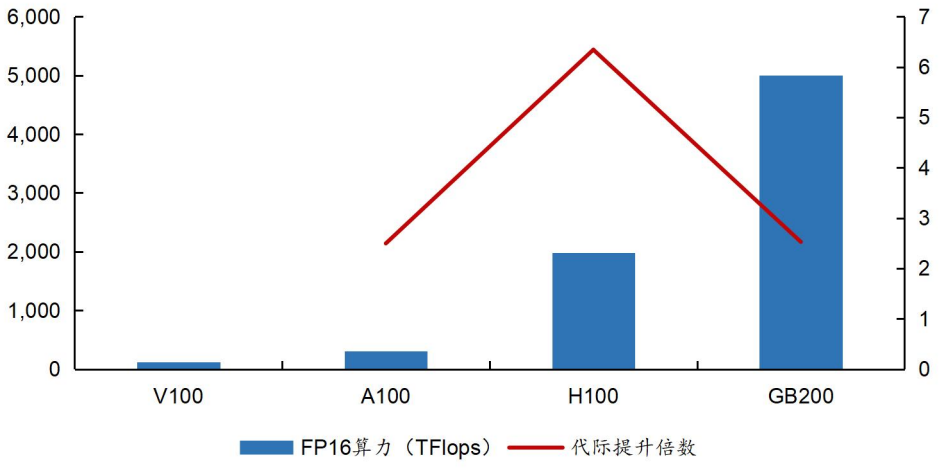
图表 13：国内外各厂商算力芯片参数对比

厂商	名称	发布时间	制程(nm)	晶体管数量(亿)	算力指标（单位：TFLOPS）						存储和互连指标			
					FP64	FP64矩阵	FP32	FP32矩阵	FP16矩阵/BF16	INT8	HBM/显存	显存带宽(GB/s)	显存容量(GB)	互联速度(GB/s)
英伟达	GB200	2023/3/19	4	2080	/	90.0	/	/	5000.0	10000.0	HBM3e	16384	384	/
英伟达	H200	2023/11/13	4	800	34.0	67.0	67.0	989.0	1979.0	3958.0	HBM3e	4915	141	900
英伟达	H100	2023/3/22	4	800	34.0	67.0	67.0	989.0	1979.0	3958.0	HBM3	3430	80	900
英伟达	H800	/	4	/	1.0	1.0	67.0	989.0	1979.0	3958.0	HBM3	3430	80	400
英伟达	A100	2020/7/10	7	542	9.7	19.5	19.5	156.0	312.0	624.0	HBM2e	2039	80	600
英伟达	A800	/	7	/	9.7	19.5	19.5	156.0	312.0	624.0	HBM2	2039	40	400
英伟达	V100	2017/5/11	12	211	7.8	7.8	15.7	15.7	125.0	62.0	HBM2	900	16	300
英伟达	L40S	2023/8/12	/	/	/	/	91.6	183.0	362.0	733.0	GDDR6	864	48	/
英伟达	L40	2023/3/22	/	/	/	/	90.5	90.5	181.1	362.0	GDDR6	864	48	/
英伟达	H20	/	/	/	1.0	/	44.0	74.0	148.0	296.0	HBM3	4096	96	/
英伟达	L20	/	/	/	/	/	59.8	59.8	119.5	239.0	GDDR6	864	48	/
英伟达	L2	/	/	/	/	/	24.1	48.3	96.5	193.0	GDDR6	300	24	/
AMD	MI300X	2023/12/6	5+6	1,530	81.7	163.4	163.4	163.4	1300.0	2600.0	HBM3	5427.2	192	896
AMD	MI300A	2023/12/6	5+6	1,460	61.3	122.6	122.6	122.6	980.6	1960.0	HBM3	5427.2	128	/
AMD	MI250	2021/11/8	6	/	45.3	90.5	90.5	/	362.1	362.1	HBM2e	3276.8	128	/
华为	昇腾910	2019/8/23	7	/	/	/	/	/	256.0	512.0	/	/	/	/
寒武纪	MLU370-X8	/	7	/	/	/	24.0	/	96.0	256.0	LPDDR5	614.4	48	200

来源：英伟达官网，AMD 官网，寒武纪官网，新智元，量子位，硬件世界，AI 科技评论等，华福证券研究所

若仅考虑英伟达 FP16 算力，代际提升速度依然很快。英伟达 A100/H100/GB200 产品的 FP16 算力分别为前代产品的 2.5/6.3/2.5 倍，在数量级上持续爆发，自 2017 年至今，GB200 的 FP16 算力已达到 V100 的 40 倍。与之对比，AI 大模型参数的爆发速度相对更快，以 GPT 为例，2018 年至 2023 年，GPT 系列模型从 1 亿参数规模大幅提升至 18000 亿。**相较于 AI 大模型由 Scaling Law 驱动的参数爆发，GPU 算力增速仍亟待提升。**

图表 14：英伟达 FP16 性能代际提升情况



来源：英伟达，机器之心，华福证券研究所
注：代际提升倍数算法=新产品 FP16 性能/老产品 FP16 性能

3.2 算力利用率：来自通信、存储等多维度的综合影响

除了以上所讨论的理论峰值之外，算力利用率也影响实际算力表现。在本文图表 1 公式中，我们明确列示了算力利用率对训练存在的影响。算力利用率（MFU）是实际吞吐量与理论最大吞吐量之比。训练大语言模型并非简单的并行任务，需要在多个 GPU 之间分布模型，并且这些 GPU 需要频繁**通信**才能共同推进训练进程。通信之外，**操作符优化、数据预处理和 GPU 内存消耗等因素**，都对算力利用率（MFU）这个衡量训练效率的指标有影响。

根据 NVIDIA、Stanford University、Microsoft Research 联合发表的论文《Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM》，文中采用 A100 GPU 集群作为实验设施，实验过程中单芯片实际吞吐量如下图所示，集群吞吐量可以通过 GPU 用量×单芯片实际吞吐量计算得到。已知 A100 峰值 FP16 吞吐量达到 312TFlops，下图中“Achieved teraFLOP/s per GPU”可以理解为“theoretical peak FLOP/s”与算力利用率的乘积，从结果来看，本次实验中 AI 训练的普遍算力利用率基本处于 44%-52%的区间。该数据与我们在图表 2 和图表 3 中统计得到的部分大模型算力利用率情况基本相仿。

图表 15：AI 训练实验数据中反映的算力利用率情况（例 1）

Number of parameters (billion)	Attention heads	Hidden size	Number of layers	Tensor model-parallel size	Pipeline model-parallel size	Number of GPUs	Batch size	Achieved teraFlOP/s per GPU	Percentage of theoretical peak FLOP/s	Achieved aggregate petaFLOP/s
1.7	24	2304	24	1	1	32	512	137	44%	4.4
3.6	32	3072	30	2	1	64	512	138	44%	8.8
7.5	32	4096	36	4	1	128	512	142	46%	18.2
18.4	48	6144	40	8	1	256	1024	135	43%	34.6
39.1	64	8192	48	8	2	512	1536	138	44%	70.8
76.1	80	10240	60	8	4	1024	1792	140	45%	143.8
145.6	96	12288	80	8	8	1536	2304	148	47%	227.1
310.1	128	16384	96	8	16	1920	2160	155	50%	297.4
529.6	128	20480	105	8	35	2520	2520	163	52%	410.2
1008.0	160	25600	128	8	64	3072	3072	163	52%	502.0

来源：NVIDIA&Stanford University&Microsoft Research《Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM》，华福证券研究所

另外，根据字节、北大联合发表的《MegaScale: Scaling Large Language Model Training to More Than 10,000 GPUs》，字节提出了一个名为 MegaScale 的生产系统，旨在解决在万卡集群上训练大模型时面临的效率和稳定性挑战。在 12288 块 GPU 上训练 1750 亿参数大语言模型时，MegaScale 实现了 55.2%的算力利用率（MFU），是英伟达 Megatron-LM 的 1.34 倍。从结果来看，本次实验中 AI 训练的普遍算力利用率基本处于 40%-66%的区间，比前述论文对比，已经有了较大的提升。

图表 16：AI 训练实验数据中反映的算力利用率情况（例 2）

Batch Size	Method	GPUs	Iteration Time (s)	Throughput (tokens/s)	Training Time (days)	MFU	Aggregate PFlops/s
768	Megatron-LM	256	40.0	39.3k	88.35	53.0%	43.3
		512	21.2	74.1k	46.86	49.9%	77.6
		768	15.2	103.8k	33.45	46.7%	111.9
		1024	11.9	132.7k	26.17	44.7%	131.9
	MegaScale	256	32.0	49.0k	70.86	65.3%(**1.23**×)	52.2
		512	16.5	95.1k	36.51	63.5%(**1.27**×)	101.4
		768	11.5	136.7k	25.40	61.3%(**1.31**×)	146.9
		1024	8.9	176.9k	19.62	59.0%(**1.32**×)	188.5
6144	Megatron-LM	3072	29.02	433.6k	8.01	48.7%	466.8
		6144	14.78	851.6k	4.08	47.8%	916.3
		8192	12.24	1027.9k	3.38	43.3%	1106.7
		12288	8.57	1466.8k	2.37	41.2%	1579.5
	MegaScale	3072	23.66	531.9k	6.53	59.1%(**1.21**×)	566.5
		6144	12.21	1030.9k	3.37	57.3%(**1.19**×)	1098.4
		8192	9.56	1315.6k	2.64	54.9%(**1.26**×)	1400.6
		12288	6.34	1984.0k	**1.75**	55.2%(**1.34**×)	2166.3

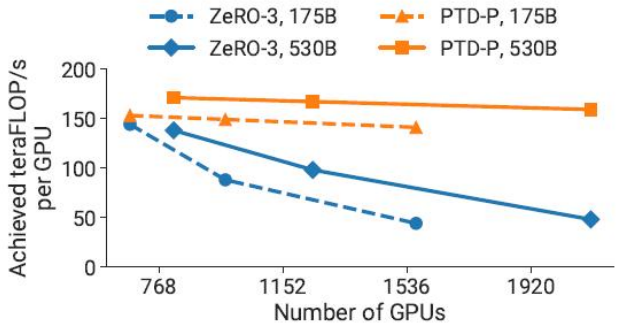
来源：ByteDance&Peking University《MegaScale: Scaling Large Language Model Training to More Than 10,000 GPUs》，华福证券研究所

我们通过以上两篇论文给出的训练结果对比大致可以了解到，在控制其他条件不变的前提下：

（1）大模型的改进对算力利用率有较大提升。字节 MegaScale 是在英伟达 Megatron-LM 的基础上改进的。具体改进包括，算法和系统组件的共同设计、通信和计算重叠的优化、操作符优化、数据流水线优化以及网络性能调优等。如图表 18 所示，MegaScale 在不同情形下的算力利用率（MFU）均显著高于 Megatron-LM。

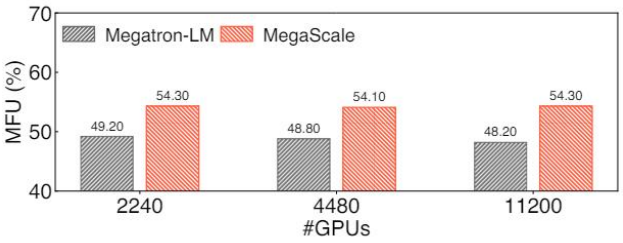
（2）大模型参数量越多，算力利用率越高。如图表 17 所示，PTD-P 和 ZeRO-3 模型在 530B 参数体量下每 GPU 实际达到的算力均高于 175B 对应算力。

图表 17：PTD-P 和 ZeRO-3 模型的单芯片吞吐量情况



来源：NVIDIA&Stanford University&Microsoft Research《Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM》，华福证券研究所

图表 18：530B 参数的 Megatron-LM 和 MegaScale 模型的算力利用率（MFU）情况



来源：ByteDance&Peking University《MegaScale: Scaling Large Language Model Training to More Than 10,000 GPUs》，华福证券研究所

4 文本大模型 AI 训练侧对 GPU 的需求量如何求解？

本文按第一章所示“文本大模型 AI 训练侧算力供给需求公式”，逐步拆解计算过程。首先，测算 AI 大模型所需要的计算量，随后通过单 GPU 算力供给能力、算力利用率等数值的假设，逐步倒推得到 GPU 需求数量。

一、AI 大模型训练侧需求

1、每参数每 token 所需计算量：我们参考 OpenAI 在 2020 年提出的 Scaling Law，论文指出：对于 Decoder-only 的模型，计算量 C、模型参数量 N、数据规模 D 三者满足 C ≈ 6ND，即每参数每 token 所需计算量为 6 Flops。

2、大模型参数量*token 数：首先，我们将大模型划分为三个梯队，可以理解为以 GPT 为代表的“第一梯队”（23 年发布的 GPT-4 已达到万亿参数水平），行业内其他知名大模型作为“第二梯队”。根据 HAI《2024 AI Index Report》，2023 年全球知名大模型共 89 个，我们将其定义为“第一/二梯队”大模型数量之和。我们对于 2024-2026 年大模型数量给出如下预测：假设随着 AI 训练所需计算量持续爆发，训练成本持续提升，叠加行业竞争加剧，我们预计第一梯队数量或将有所减少；假设大模型从通用向垂直行业延伸的趋势持续演绎，第二梯队及其他大模型数量或将持续上升。此外，我们认为 Scaling Law 仍将持续存在，各梯队大模型参数或将持续通过提升参数量、预训练数据规模（token 数）带动计算量提升，进而提升大模型性能，按过往提升速度大致推断未来增长情况，而参数量与预训练数据规模（token 数）的关系参考 Chinchilla 法则进行假设预测。

二、AI 大模型训练侧供给

1、GPU 计算性能：鉴于英伟达当前在 AI 训练卡方面的龙头地位以及长期以来较高的市占率情况，我们以英伟达训练卡性能来进行粗略估算。我们基于图表 14 中的数据，英伟达 V100 至 GB200 在 FP16 算力代际提升方面，新产品分别为前代产品的 2.5/6.3/2.5 倍。由于 AI 训练端对于数据精度的要求一般为 FP16/INT8，暂时不考虑 FP4 架构的大规模使用。我们假设未来英伟达新产品的 FP16 算力在 Blackwell 架构的基础上延续过往倍增趋势。此外，从英伟达已有产品的实际产能情况看，大部分产品实际应用到产业界会比发布时间有较大的延迟，主要由于产能方面的紧缺性。据 Digitimes，23H2 英伟达 H100 GPU 的交货时间达到 11 个月，进入 2024 年以来，交货时间显著缩短，逐步缩短到仅 2-3 个月（8-12 周）。由此我们假设，2023/2024 年分别以 Ampere/Hopper 为主，后续逐步升级迭代。

2、训练时间&算力利用率：我们假设 AI 大模型训练卡长期供不应求，全年算力设施接近满负荷运转，由此假设全年有效训练时间为 350 天。我们参考 GPT 大模型的算力利用率情况，我们认为 GPT 与英伟达 GPU 的适配程度以及训练效率或为行业前沿水平，因此我们假设行业一般水平在未来几年维持在 30-42%区间。

三、结论：以英伟达 Hopper/Blackwell/下一代 GPU 卡 FP16 算力衡量，我们认为 2024-2026 年全球文本大模型 AI 训练侧 GPU 需求量为 271/592/1244 万张。

图表 19：全球文本大模型 AI 训练侧算力需求-供给测算

市场空间测算｜需求侧							
假设							
1、我们将GPT量级为代表的大模型作为“第一梯队”，将行业内其他知名大模型作为“第二梯队”							
2、根据HAI《2024 AI Index Report》，2023年全球知名大模型共89个，我们将其定义为“第一/二梯队”大模型数量之和							
3、假设随着AI训练所需计算量持续爆发，训练成本持续提升，叠加行业竞争加剧，我们预计第一梯队数量或将有所减少							
4、假设大模型从通用向垂直行业延伸的趋势持续演绎，第二梯队及其他大模型数量或将持续上升							
5、根据Scaling Law和Chinchilla法则，我们给出大模型参数量与预训练数据规模（tokens）预测情况如下：							
				2023	**2024E**	**2025E**	**2026E**
全球文本大模型AI训练侧算力需求		**ZFlops**	**10^21Flops**	**4465152**	**51910800**	**322183200**	**1974210000**
每参数每token所需计算量				6	6	6	6
单个大模型参数量*token数							
第一梯队		ZFlops	10^21Flops	23400	245000	980000	2205000
	参数量		万亿	1.8	3.5	7.0	10.5
	token数		万亿	13	70	140	210
第二梯队				7200	23400	245000	980000
	参数量			0.6	1.8	3.5	7.0
	token数			12	13	70	140
其他				72	7200	23400	245000
	参数量			0.1	0.6	1.8	3.5
	token数			1	12	13	70
大模型数			个	400	753	938	998
第一梯队				5	5	4	3
第二梯队				84	126	126	107
其他				311	622	808	888
市场空间测算｜供给侧							
假设							
1、假设全年算力设施接近满负荷运转							
2、假设以当年英伟达主要出货的GPU产品作为计算基准							
3、已知英伟达H100到GB200，FP16算力后代为前代的2.5x，我们假设下一代产品算力也为GB200的2.5x							
4、假设GPT系列大模型的算力利用率为行业领先水平，我们假设行业平均水平相对低于GPT系列							
5、假设随着AI大模型行业发展，AI训练侧算力利用率逐年提升							
				2023	**2024E**	**2025E**	**2026E**
全球文本大模型AI训练侧GPU需求量			**万张**	**158**	**271**	**592**	**1244**
每秒平均算力需求		ZFlops	10^21Flops	0.5	5.4	29.6	155.4
	算力利用率			30%	32%	36%	42%
	年有效训练时长		天	350	350	350	350
当年英伟达主要出货GPU				Ampere	Hopper	Blackwell	下一代产品
FP16计算性能		TFlops		312	1979	5000	12500

来源： NVIDIA&Stanford University&Microsoft Research《Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM》，OpenAI《Scaling Laws for Neural Language Models》，HAI《2024 AI Index Report》，英伟达，Digitimes，华尔街见闻等，华福证券研究所测算

我们进一步将此结果与市场一致预期作对比：根据腾讯科技援引 Digitimes 预测，台积电 24Q1 CoWoS 产能将爬升到 17000 片/月，到年底有机会爬升到 26000 片-28000 片/月。按照 CoWoS 月产能 17000 片的数据来计算，如果英伟达可以拿到其中 40%，即 6800 片，而一片 12 英寸的晶圆，大致可以切 30 张左右的 H200，即台积电单月可完成 20.4 万张 H200 的封装。到年底，按照台积电 26000 片/月的 CoWoS 产能，英伟达如果还是占 40%，即 10400 片/月，单月可以完成 31.2 万张 H200 的封装。也即，英伟达在台积电的助攻下，**H200 GPU 全年的封装产能，下限可能是 244 万张，上限有可能突破 374 万张。**

我们对于2024年的预期略低于台积电CoWoS产能指引GPU出货量预期均值的原因主要有以下几点：

1、本文对于2024年算力供给的测算，有一个前提假设是“以当年英伟达主要出货的GPU产品作为计算基准”,也即我们得到的训练卡需求是271万张Hopper GPU对应的算力规模，而实际应用中，由于GPU供不应求的现象持续存在，加之Hopper GPU产能在2024年才逐步得以爬坡上量，年初部分大模型训练或仍采用Ampere GPU或其他略低端的训练卡。

2、本文仅限于对文本大模型AI训练侧需求的测算，尚未包括Sora等视频生成类大模型对应的算力需求，建议持续关注多模态趋势对AI训练侧GPU需求的驱动作用。

5 风险提示

AI需求不及预期风险、Scaling Law失效风险、GPU技术升级不及预期的风险、测算模型假设存在偏差风险。

分析师声明

本人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，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，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收到任何形式的补偿。

一般声明

华福证券有限责任公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。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，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告的信息均来源于本公司认为可信的公开资料，该等公开资料的准确性及完整性由其发布者负责，本公司及其研究人员对该等信息不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，之后可能会随情况的变化而调整。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。本公司不保证本报告所含信息及资料保持在最新状态，对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

在任何情况下，本报告所载的信息或所做出的任何建议、意见及推测并不构成所述证券买卖的出价或询价，也不构成对所述金融产品、产品发行或管理人作出任何形式的保证。在任何情况下，本公司仅承诺以勤勉的职业态度，独立、客观地出具本报告以供投资者参考，但不就本报告中的任何内容对任何投资做出任何形式的承诺或担保。投资者应自行决策，自担投资风险。

特别声明

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。投资者请勿将本报告视为投资或其他决定的唯一参考依据。

投资评级声明

类别	评级	评级说明
公司评级	买入	未来 6 个月内，个股相对市场基准指数指数涨幅在 20%以上
	持有	未来 6 个月内，个股相对市场基准指数指数涨幅介于 10%与 20%之间
	中性	未来 6 个月内，个股相对市场基准指数涨幅介于-10%与 10%之间
	回避	未来 6 个月内，个股相对市场基准指数涨幅介于-20%与-10%之间
	卖出	未来 6 个月内，个股相对市场基准指数涨幅在-20%以下
行业评级	强于大市	未来 6 个月内，行业整体回报高于市场基准指数 5%以上
	跟随大市	未来 6 个月内，行业整体回报介于市场基准指数-5%与 5%之间
	弱于大市	未来 6 个月内，行业整体回报低于市场基准指数-5%以下

备注：评级标准为报告发布日后的 6~12 个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中 A 股市场以沪深 300 指数为基准；香港市场以恒生指数为基准，美股市场以标普 500 指数或纳斯达克综合指数为基准（另有说明的除外）

联系方式

华福证券研究所 上海
公司地址：上海市浦东新区浦明路 1436 号陆家嘴滨江中心 MT 座 20 层
邮编：200120
邮箱：hfyjs@hfzq.com.cn