英伟达（NVDA.US）

加速，规模，超线性

评级：	增持
当前价格（美元）：	495.22

2024.01.03

秦和平(分析师)
0755-23976666
qinheping027734@gtjas.com
证书编号 S0880523110003

交易数据

52 周内股价区间（美元）	140.36-504.09
当前股本（百万股）	2,487
当前市值（百万美元）	1.221

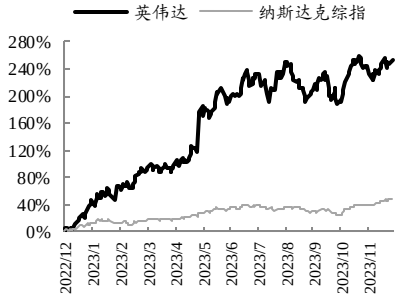
本报告导读：

AIGC 催生巨大的加速计算需求，从通用计算向加速计算转型初期，长期扩大的加速计算需求助推系统性领先的平台型计算公司英伟达进入长期超线性增长。

摘要：

- **盈利预测与投资建议：**我们认为英伟达数据中心业务呈现超线性增长，因此我们调高英伟达 FY2024E-FY2026E 的营业收入分别为 599/ 966 / 1,190 亿美元（前值为 599/891/1004 亿美元），同增 122%/ 61%/ 23%。由于毛利率提升、费用率缩小，使得经调整净利润在未来 3 年呈现大幅提升，我们调高英伟达 FY2024E-FY2026E 的经调整净利润为 313/541/665 亿美元（前值为 313/500/562 亿美元）。我们使用 DCF 和 PE BAND 估值法，并综合两种估值方法的结果，上调英伟达 FY2025 目标价至 633 USD，维持“增持”评级。
- **超预期：**英伟达以其 GPU、GraceCPU、BluefieldDPU 组成的超异构计算技术演进系统，以及 CUDA 编程平台构建的平台型系统竞争优势，将在通用计算向加速转型过程中进入长期超线性增长。**市场认为，**AIGC 技术变革存在不确定性，且目前预期的英伟达的增长空间主要依赖大模型训练和推理 GPU 算力需求，短期线性比例外推英伟达的收入增长；**我们认为**，AIGC 技术趋势已经确立，英伟达凭借其领先的平台型系统竞争迭代优势，将获得长期超线性增长。
- **核心信息与逻辑：**这种超线性增长主要来源于 1）异构计算系统及 CUDA 技术平台领先且快速迭代带来的规模及成本优势；2）因围绕 GPU 加速系统所形成的新型技术生态网络设备 infiniteband 网络设备系统的增长；3）应用，行业，区域的加速计算技术扩散。
- **催化剂：**1）台积电 CoWoS 等芯片供应链产能提升到足够供给；2）大模型准确度，开源性，终端化更成熟；3）应用场景和区域渗透快速落地。
- **风险提示：**大模型技术成熟速度及安全规范控制不及预期；地缘政治限制芯片自由贸易；技术行业可能会受到宏观经济周期的影响。在经济衰退期间，技术产品和服务的需求可能会下降；竞争风险等。

52周内股价走势图



相关报告

《全球计算转型初期，世界 AI 引擎全力加速—英伟达 3QFY24 业绩点评》 2023.11.22

《英伟达持续狂奔，看好 AI 趋势的能见度至 2024 年》 2023.11.14

《算力需求激增，全球 AI 引擎加速—英伟达 2QFY24 业绩点评》 2023.08.25

《AI 算力新供给：AMD MI 300 表现亮眼，英伟达市场地位稳固》 2023.06.15

《重塑计算，世界 AI 的引擎—英伟达首次覆盖报告》 2023.06.05

财务摘要（百万美元）	FY2020A	FY2021A	FY2022A	FY2023A	FY2024E	FY2025E	FY2026E
营业收入	10,918	16,675	26,914	26,974	59,894	96,623	119,022
(+/-)%	-	52.7%	61.4%	0.2%	122.0%	61.3%	23.2%
毛利润	6,768	10,396	17,475	15,356	43,211	71,984	88,671
(+/-)%	-	53.6%	68.1%	-12.1%	181.4%	66.6%	23.2%
经调整净利润	3,580	6,277	11,259	8,365	31,326	54,115	66,544
(+/-)%	-	75.3%	79.4%	-25.7%	274.5%	72.8%	23.0%
经调整 PE	1,367.80	198.02	111.50	148.42	39.43	22.82	18.56

目 录

1. 投资建议

核心数据预测：我们认为英伟达数据中心业务呈现超线性增长，因此我们调高英伟达 FY2024E-FY2026E 的营业收入分别为 599/ 966 / 1,190 亿美元（前值为 599/891/1004 亿美元），同增 122%/ 61%/ 23%。由于毛利率提升、费用率缩小，使得经调整净利润在未来 3 年呈现大幅提升，我们调高英伟达 FY2024E-FY2026E 的经调整净利润为 313/541/665 亿美元（前值为 313/500/562 亿美元）。

核心假设：

1）营收主要由数据中心业务驱动，数据中心由于 AI 产业的高成长性呈现超线性增长；

2）公司毛利率在 2024 财年由于数据中心业务的高毛利率增长至 72%，并于 2026 财年稳定至 75% 左右；

3）费用率，包含营销费用率、一般及行政费用率和研发开支率，由于规模效应的释放与企业经营杠杆的增强大幅缩小，三年维度稳定至 13% 左右；

4）净利润率由于毛利率的提升、费用率的缩小，在 2024 年呈现大幅提升，随及在未来两年趋于稳定，我们认为净利润率将稳定至 51% 左右。

图 1 英伟达核心指标预测（单位：USD MN）

USD MN		FY2023	FY2024E	FY2025E	FY2026E	3QFY24	4QFY24E	1QFY25E	2QFY25E	3QFY25E
Period End		2023/1/29	2024/1/31	2025/1/31	2026/1/31	2023/10/29	2024/1/31	2024/4/30	2024/7/31	2024/10/31
营收(Revenue)	mn	26,974	59,894	96,623	119,022	18,120	21,075	21,470	22,931	24,505
同比增长(% yoy growth)	%	*0.2%*	*122.0%*	*61.3%*	*23.2%*	*205.5%*	*248.3%*	*198.5%*	*69.8%*	*35.2%*
拆分By Segment：										
数据中心(Data Center)	mn	15,005	46,297	80,169	100,211	14,514	17,176	17,564	19,098	20,320
同比增长(%yoy growth)	%	*41%*	*209%*	*73%*	*25%*	*279%*	*375%*	*310%*	*85%*	*40%*
游戏(Gaming)	mn	9,067	10,603	12,741	14,361	2,856	3,021	3,024	2,958	3,284
同比增长(%yoy growth)	%	*-27%*	*17%*	*20%*	*13%*	*81%*	*65%*	*35%*	*19%*	*15%*
专业可视化(Professional Visualization)	mn	1,544	1,510	1,829	1,974	416	420	413	459	483
同比增长(%yoy growth)	%	*-27%*	*-2%*	*21%*	*8%*	*108%*	*86%*	*40%*	*21%*	*16%*
自动驾驶(Automotive)	mn	903	1,154	1,500	2,074	261	344	358	339	342
同比增长(%yoy growth)	%	*60%*	*28%*	*30%*	*38%*	*4%*	*17%*	*21%*	*34%*	*31%*
OEM及IP(OEM &IP)	mn	455	329	383	401	73	113	111	77	77
同比增长(%yoy growth)	%	*-61%*	*-28%*	*17%*	*5%*	*0%*	*35%*	*44%*	*17%*	*5%*
毛利（Gross Profit）	mn	15,356	43,211	71,984	88,671	13,400	15,701	15,995	17,083	18,256
毛利率(% Gross Profit Margin)	%	*56.9%*	*72.1%*	*74.5%*	*74.5%*	*74.0%*	*74.5%*	*74.5%*	*74.5%*	*74.5%*
费用（Total Opex）	mn	11,131	11,398	14,880	15,657	2,983	3,245	3,306	3,531	3,774
费用率Total Opex %	%	*41.3%*	*19.0%*	*15.4%*	*13.2%*	*16.5%*	*15.4%*	*15.4%*	*15.4%*	*15.4%*
S&M + G&A	mn	2,440	2,661	3,285	4,047	689	717	730	780	833
S&M % + G&A %	%	*9.0%*	*4.4%*	*3.4%*	*3.4%*	*3.8%*	*3.4%*	*3.4%*	*3.4%*	*3.4%*
研发开支(R&D)	mn	7,338	8,738	11,595	14,283	2,294	2,529	2,576	2,752	2,941
研发开支(R&D) %	%	*27.2%*	*14.6%*	*12.0%*	*12.0%*	*12.7%*	*12.0%*	*12.0%*	*12.0%*	*12.0%*
经营利润（Ebit, Operating Income）	mn	4,225	31,812	57,104	73,014	10,417	12,455	12,689	13,552	14,483
经营利润率(% Op Margin)	%	*15.7%*	*53.1%*	*59.1%*	*61.3%*	*57.5%*	*59.1%*	*59.1%*	*59.1%*	*59.1%*
经调整经营利润（Ebit, Non-GAAP Operaing Income）	mn	9,039	35,936	62,128	78,474	11,557	13,551	13,805	14,744	15,757
经营利润率(% Op Margin)	%	*33.5%*	*60.0%*	*64.3%*	*65.9%*	*63.8%*	*64.3%*	*64.3%*	*64.3%*	*64.3%*
净利润（GAAP,Net Income）	mn	4,368	28,231	49,054	60,314	9,243	10,757	10,904	11,652	12,438
净利润率(% Net Profit Margin)	%	*16.2%*	*47.1%*	*50.8%*	*50.7%*	*51.0%*	*51.0%*	*50.8%*	*50.8%*	*50.8%*
经调整净利润（Non-GAAP Net Profit）	mn	8,365	31,326	54,115	66,544	10,020	11,853	12,026	12,848	13,732
非GAAP的净利润率(% Non-GAAP Net Profit Margin)	%	*31.0%*	*52.3%*	*56.0%*	*55.9%*	*55.3%*	*56.2%*	*56.0%*	*56.0%*	*56.0%*
摊薄GAAP每股收益（Net profit per ADS, diluted, GAAP）	mn	1.74	11.32	19.67	24.18	3.71	4.31	4.37	4.67	4.99
摊薄Non-GAAP每股收益（Net profit per share, diluted, NON-GAAP）	mn	3.34	12.56	21.70	26.68	4.02	4.75	4.82	5.15	5.51

数据来源：英伟达财报，国泰君安证券研究预测

估值 1：考虑到英伟达即将在 2024 财年产生持续稳定的利润和现金流，我们首先采用 DCF 估值法对公司的股权价值进行测算。基于以下

假设，我们计算得到公司 2025 财年（对应日历年 2024/01/31-2025/01/31）的公司估值为 1.59 万亿美金，对应每股价格为 637 USD。

核心假设：

1）采用美国十年期国债收益率为无风险收益率 Rf=4.3%；

2）风险溢价：我们计算得出 2013 年到 2023 年道琼斯指数的复合增长率为 8.6%，得出风险溢价为 4.3%；

3）β=1.69；由于英伟达没有债权成本，基于核心假设 1 和核心假设 2 我们得出加权平均资金成本 WACC 为 11.5%；

4）永续增长率为 2.8%。

图 2 英伟达 DCF 估值法

USD MN Period End		FY2020 2020/1/26	FY2021 2021/1/31	FY2022 2022/1/30	FY2023 2023/1/29	FY2024E 2024/1/31	FY2025E 2025/1/31	FY2026E 2026/1/31	FY2027E 2027/1/31	FY2028E 2028/1/31	FY2029E 2029/1/31	FY2030E 2030/1/31	FY2031E 2030/2/1	FY2032E 2030/2/2	FY2033E 2030/2/3	FY2034E 2030/2/4
营收(Revenue)	mn	10,918	16,675	26,914	26,974	59,894	96,623	119,022	142,826	168,535	197,186	228,735	263,046	302,502	347,878	400,059
同比增长(% yoy growth)	%	-7%	53%	61%	0%	122%	61%	23%	20%	18%	17%	16%	15%	15%	15%	15%
毛利(Gross Profit)	mn	6,768	10,396	17,474	15,356	43,211	71,984	88,671	107,119	126,907	148,875	172,695	199,915	229,902	267,866	308,046
毛利率(% Gross Profit Margin)	%	62.0%	62.3%	64.9%	56.9%	72.1%	74.5%	74.5%	75.0%	75.3%	75.5%	75.5%	76.0%	76.0%	77.0%	77.0%
经调整经营利润(Ebit, Non-GAAP Operaing Income)	mn	3,736	6,804	12,691	9,039	35,936	62,128	78,474	95,693	114,941	136,058	160,115	185,447	214,777	246,993	284,042
经营利润率(% Op Margin)	%	34.2%	40.8%	47.2%	33.5%	60.0%	64.3%	65.9%	67.0%	68.2%	69.0%	70.0%	70.5%	71.0%	71.0%	71.0%
+ 税 (Tax)	mn	175	76	188	(186)	4,136	8,657	10,644	14,354	17,241	20,409	24,017	27,817	32,217	37,049	42,606
+ 折旧与摊销(D&A)	mn	381	1,098	1,174	1,544	3,414	5,508	6,784	8,141	9,606	11,240	13,038	14,994	17,243	19,829	22,803
- 支出 Capex	mn	489	1,128	976	1,833	3,209	5,222	6,653	7,855	9,269	10,845	12,580	14,468	16,638	19,133	22,003
+ 运营资本的变化 Increase In Working Capital	mn	717	(703)	(3,363)	(2,207)	(3,801)	(4,420)	(3,517)	(3,462)	(3,481)	(3,736)	(3,723)	(3,584)	(3,597)	(3,624)	(3,653)
自由现金流 (FCF)	mn	4,170	5,995	9,338	6,729	28,204	49,338	64,445	78,163	94,555	112,308	132,832	154,572	179,568	207,016	238,583
Terminal Value	mn															2,809,109
Total	mn	4,170	5,995	9,338	6,729	28,204	49,338	64,445	78,163	94,555	112,308	132,832	154,572	179,568	207,016	2,809,109

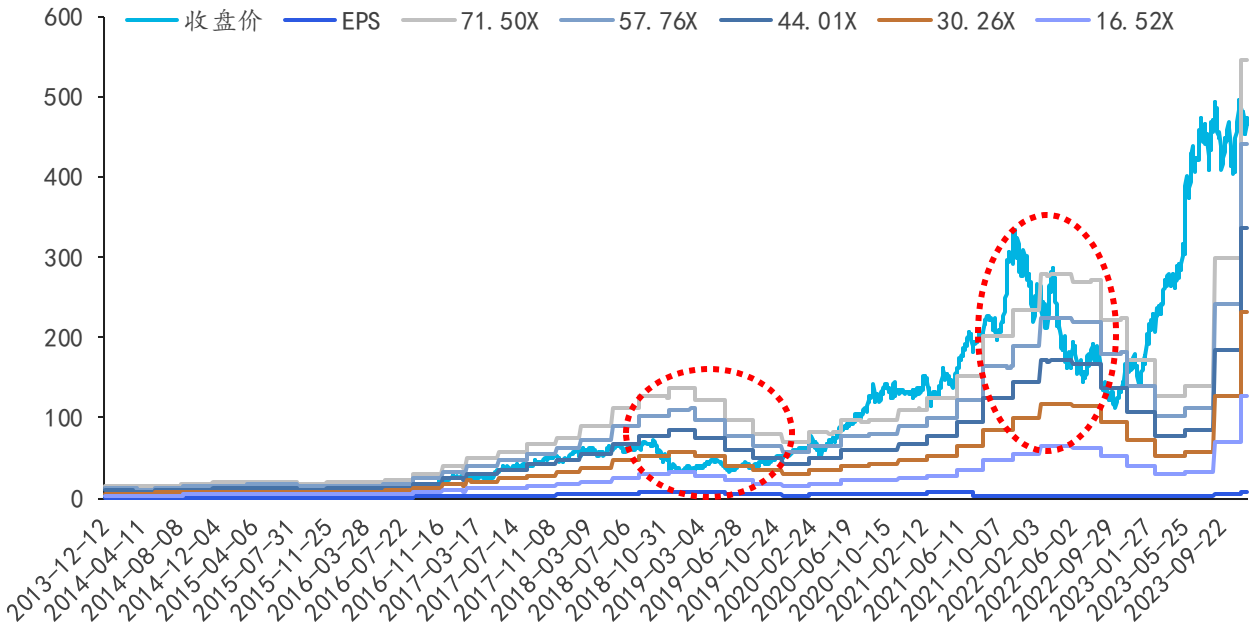
DCF 核心假设		
无风险利率 Rf	%	4.3%
风险溢价 Equity Market Premium	%	4.3%
Beta β		1.69
加权平均资金成本 WACC	%	11.5%
永续增长率 Perpetuity Growth Rate	%	2.8%
目标价		FY2024E / FY2025E
企业价值 Enterprise Value	mn	1,383,954 / 1,515,333
现金 Net Cash	mn	27,851 / 73,694
目标市值 Equity Value(USD)	mn	1,411,804 / 1,589,027
目标股价 Target Price	USD	**566.1** / **637.1**

数据来源：英伟达财报，国泰君安证券研究预测

估值 2：我们认为英伟达在 AI 算力领域持续狂奔，不断突破原有产品的能力本身是英伟达的核心竞争力，强大的竞争壁垒可以实现公司在 AI 算力行业中保持强大的盈利能力和持久性，其在数据中心高端 GPU 方面的技术和市场优势极为明显，短中长期内难以撼动，存在极强的标的稀缺性。

英伟达是典型的周期股特征。我们类比英伟达在加密货币浪潮，与疫情下用户对电脑游戏的需求两次周期下的估值中枢，预计英伟达 GPU 的需求将在 2025 财年延续强劲的需求，且 2025 财年每个季度的供应量都会增加，处于上涨周期。从过去 10 年的 PE-BAND 以及略高于中位数的可比公司 2024 年 PE28X 的估值估计，我们给予英伟达 FY2025E PE 32X，对应公司估值为 1.57 万亿美金，对应每股价格为 629USD。

图 3 2013 年-2023 年英伟达股价变化与 PE Bands



数据来源：iFinD，国泰君安证券研究

图 4 英伟达可比公司估值

代码	名称	股价（美元）	EPS			PE		
			FY2023	FY2024E	FY2025E	FY2023	FY2024E	FY2025E
NVDA.US	英伟达	494.17	1.76	12.30	20.46	83.31	40.19	24.16
中位数			3.36	4.59	4.92	30.16	27.85	22.51
平均值			6.54	8.67	10.23	36.15	28.46	23.34
AMD.US	AMD	146.07	2.65	3.80	5.04	55.02	38.48	29.01
INTC.US	英特尔	50.76	0.95	1.84	2.54	53.38	27.61	20.02
AVGO.US	博通股份	1126.17	33.93	46.98	55.86	24.80	23.97	20.16
QCOM.US	高通公司	145.72	6.47	9.24	10.62	14.51	15.77	13.73
LSCC.US	莱迪斯半导体	71.30	2.01	2.02	2.46	35.52	35.26	28.98
ADI.US	亚德诺半导体	199.35	6.60	7.10	8.76	15.92	28.08	22.76
MRVL.US	美满电子科技	61.26	-0.19	1.51	2.02	78.72	40.51	30.36
MCHP.US	微芯科技	91.12	4.07	5.37	4.81	20.70	16.96	18.95
RMBS.US	Rambus	68.49	1.76	2.12	2.52	38.91	32.28	27.16
TXN.US	德州仪器	171.23	7.13	6.67	7.69	24.02	25.69	22.27

注：数据截至 2023/12/28，各公司数据以其最新财年年报计算，NVDA、QCOM、ADI、MRVL 和 MCHP 的 FY2023 EPS 和 PE 为实际值，其余公司由于财年未结束，均为彭博一致预测值。

估值结论：综合两种估值方法，我们上调英伟达 2025 财年的目标价至 633USD，维持“增持”评级。

图 5 英伟达财务报表预测（单位：百万美元）

利润表						单位：百万美元
指标名称	FY2021	FY2022	FY2023	FY2024E	FY2025E	FY2026E
GAAP						
主营业务收入	16,675	26,914	26,974	59,894	96,623	119,022
主营业务成本	6,279	9,440	11,618	16,683	24,639	30,351
毛利	10,396	17,474	15,356	43,211	71,984	88,671
销售及营销开支+一般及行政开支	1,938	2,166	2,440	2,661	3,285	4,047
研发开支	3,926	5,267	7,338	8,738	11,595	14,283
经营利润	4,532	10,041	4,225	31,812	57,104	73,014
除税前净利润	4,408	9,941	4,182	32,367	57,710	70,958
净利润	4,332	9,753	4,368	28,231	49,054	60,314
摊薄GAAP每股收益	1.73	3.85	1.74	11.32	19.67	24.18
NONGAAP						
收入	16,675	26,914	26,974	59,894	96,623	119,022
主营业务成本	5,727	8,946	11,009	15,984	23,673	29,160
毛利	10,948	17,968	15,965	43,910	72,950	89,861
期间费用	4,144	5,277	6,926	7,974	10,822	11,387
经营利润	6,804	12,691	9,039	35,936	62,128	78,474
除税前净利润	6,684	12,493	9,045	36,515	62,772	79,131
净利润	6,277	11,259	8,365	31,326	54,115	66,544
摊薄NONGAAP每股收益	2.50	4.44	3.34	12.56	21.70	26.68

现金流量表						单位：百万美元
指标名称	FY2021	FY2022	FY2023	FY2024E	FY2025E	FY2026E
经营活动产生的现金流量	5,822	9,108	5,641	27,844	50,141	63,581
投资活动产生的现金流量	(19,675)	(9,830)	7,375	(3,209)	(5,222)	(6,653)
融资活动产生的现金流量	3,804	1,865	(11,617)	(174)	924	874
现金、现金等价物受限制资金的增加	(10,049)	1,143	1,399	24,462	45,844	57,802
年初现金、现金等价物及受限制资金	10,896	847	1,990	3,389	27,851	73,694
年末现金、现金等价物及受限制资金	847	1,990	3,389	27,851	73,694	131,497

资产负债表						单位：百万美元
指标名称	FY2021	FY2022	FY2023	FY2024E	FY2025E	FY2026E
流动资产合计	**16,055**	**28,829**	**23,073**	**55,729**	**112,885**	**178,253**
现金及现金等价物	847	1,990	3,389	27,851	73,694	131,497
有价证券	10,714	19,218	9,907	9,907	9,907	9,907
应收帐款净额	2,429	4,650	3,827	8,788	14,664	18,063
存货	1,826	2,605	5,159	7,447	11,818	15,335
预付费用和其他流动资产	239	366	791	1,737	2,802	3,452
非流动资产合计	**12,736**	**15,358**	**18,109**	**17,904**	**17,618**	**17,487**
物业及设备净额	2,149	2,778	3,807	3,602	3,316	3,185
经营租赁资产	707	829	1,038	1,038	1,038	1,038
商誉	4,193	4,349	4,372	4,372	4,372	4,372
无形资产净额	2,737	2,339	1,676	1,676	1,676	1,676
递延所得税资产	806	1,222	3,396	3,396	3,396	3,396
其他资产	2,144	3,841	3,820	3,820	3,820	3,820
流动负债合计	**3,925**	**4,335**	**6,563**	**10,783**	**18,600**	**23,522**
应付账款	1,149	1,783	1,193	3,425	5,787	7,057
应计负债和其他流动负债	1,777	2,552	4,120	6,282	10,812	13,590
短期债务	999	-	1,250	1,076	2,000	2,875
非流动负债合计	**7,973**	**13,240**	**12,518**	**12,518**	**12,518**	**12,518**
长期借款	5,964	10,946	9,703	9,703	9,703	9,703
长期经营租赁负债	634	741	902	902	902	902
其他长期负债	1,375	1,553	1,913	1,913	1,913	1,913
总资产	**28,791**	**44,187**	**41,182**	**73,633**	**130,503**	**195,740**
总负债	**11,898**	**17,575**	**19,081**	**23,301**	**31,118**	**36,040**
总权益	**16,893**	**26,612**	**22,101**	**50,332**	**99,386**	**159,700**

数据来源：国泰君安证券研究预测

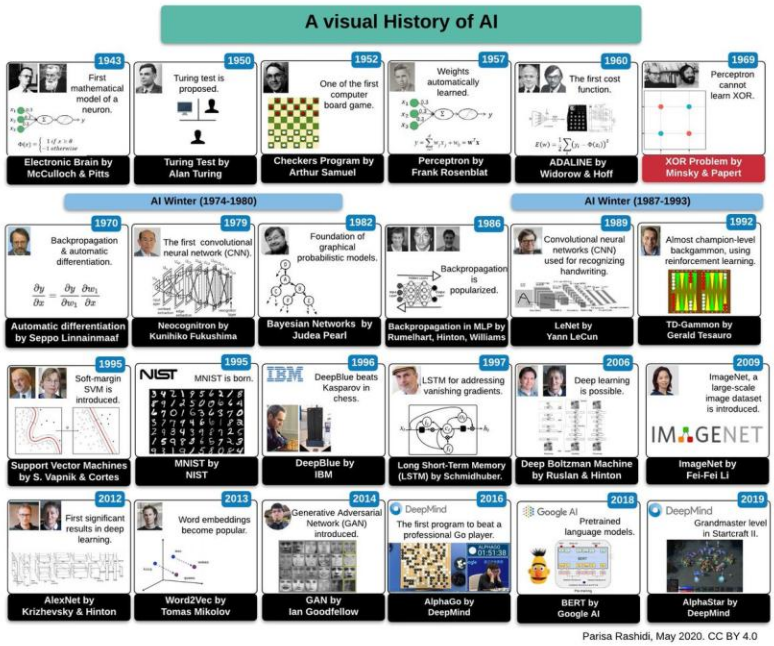
2. 生成式 AI 加速计算转型

2.1. AI-LLM

复盘 AI 的发展历史，AI 时代才刚刚开始。参考 Github 博客算法进阶的观点，AI 发展历史可分为：1）起步发展期（1943-1960 年代）：人工智能概念的提出后，发展出了符号主义、联结主义(神经网络)，相继取得了如机器定理证明、跳棋程序、人机对话等成就；2）反思发展期（1970 年代）：神经网络的训练算法、知识库、专家系统等技术理论取得了一定的革新，算力和理论依旧匮乏；3）应用发展期（1980 年代）：AI 从理论研究走向实际应用、从一般推理策略探讨转向运用专门知识的重大突破；4）平稳发展期（1990-2010 年）：伴随互联网技术狂奔，AI 创新研究加速，促使 AI 进一步走向实用化，研究的重心从基于知识系统转向了机

器学习方向，支持向量机、AdaBoost、LSTM、RNN、随机森林等算法逐步成熟；5）蓬勃发展期（2011 年至今）：随着大数据、云计算、互联网、物联网发展，基于 GPU 等的计算平台推动以深度神经网络为代表的 AI 技术飞跃，图像分类、语音识别、知识问答、人机对弈、自动驾驶等技术问世，AI 迎来爆发式增长。

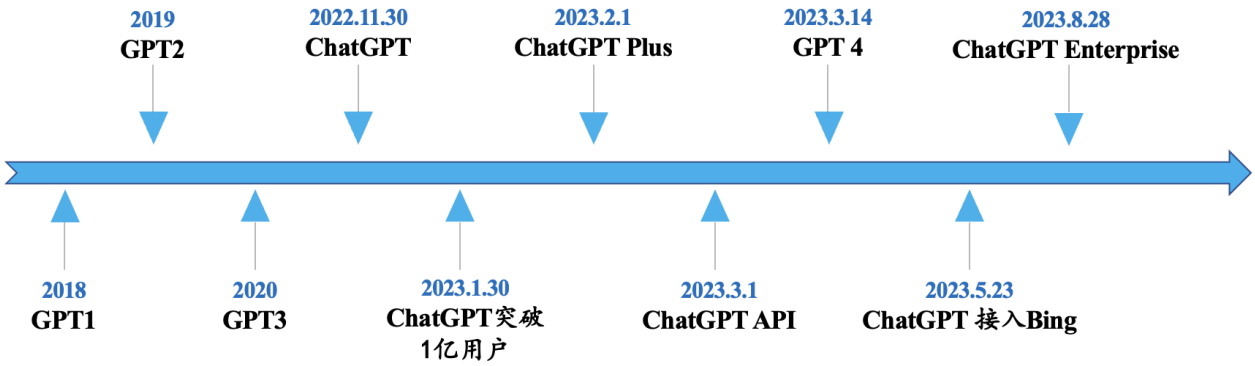
图 6 1943-2019 AI 的发展历史



数据来源：Parisa Rashidi，算法进阶

AI 进入大模型时代，ChatGPT 开创人类与 AI 交互时代。2022 年 11 月 30 日，OpenAI 开发的 Chat-GPT 横空出世，仅仅花了 5 天时间，ChatGPT 的注册用户数量达到 100 万；2023 年 3 月 14 日 GPT4 发布，GPT 正式迈向多模态。过去一年间，基于大语言模型（LLM）的生成式 AI 迅速在全球引起巨大影响。

图 7 一年来 ChatGPT 发展时间线

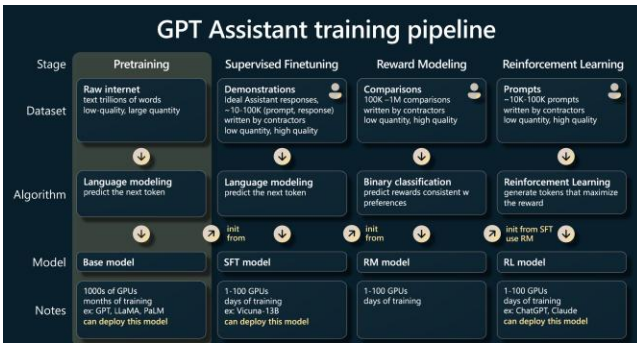


数据来源：officetimeline，国泰君安证券研究

大模型的训练步骤繁琐，使得对算力的需求激增。2023 年的微软 Build 开发者大会中，特斯拉前 AI 总监 Andrej Karpathy 发表题为 GPT 现状的

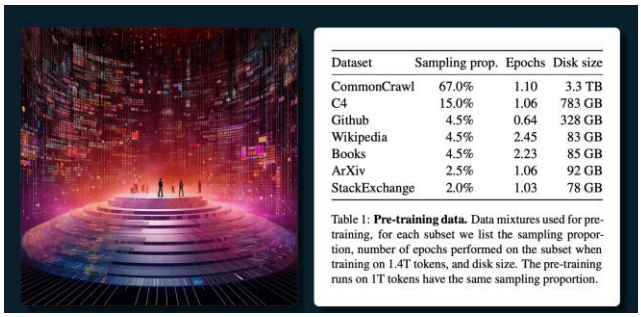
演讲。训练 ChatGPT 等大语言模型分为标记化（Tokenization）、预训练（Pretraining）、监督微调（Supervised Finetuning）和人类反馈强化学习(RLHF)四个步骤。在进行预训练之前，有 2 个准备步骤:1）数据收集:例如 Meta LLaMA 模型从 Github、维基百科等来源收集大量混合数据；2）标记化，将文本中的单词标记并转换为整数。例如 175B 参数的 GPT-3 在 300B 个 token 上训练，而 65B 参数的 LLaMA 已经在 1-1.4T 个 token 上训练，反映了模型参数的增加对于 GPU 训练需求的飙升，但参数大并不等同于模型性能强。

图 8 GPT 训练分为四个阶段



数据来源：微软 Build 大会

图 9 预训练阶段需要大量的计算资源和数据集

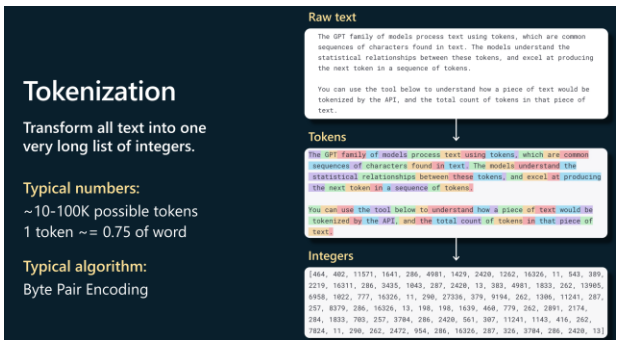


Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.

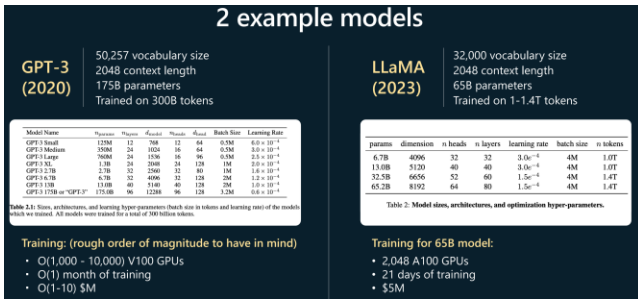
数据来源：微软 Build 大会

图 10 对文本中的单词进行标记化处理



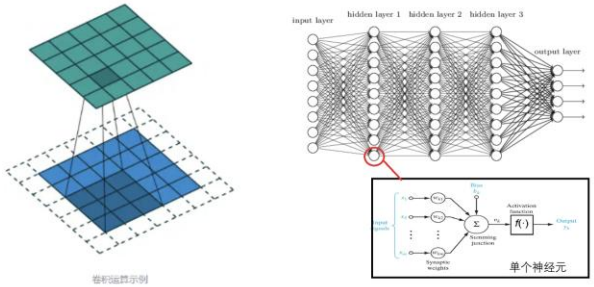
数据来源：微软 Build 大会

图 11 GPT-3 和 LLaMA 的标记化过程对比



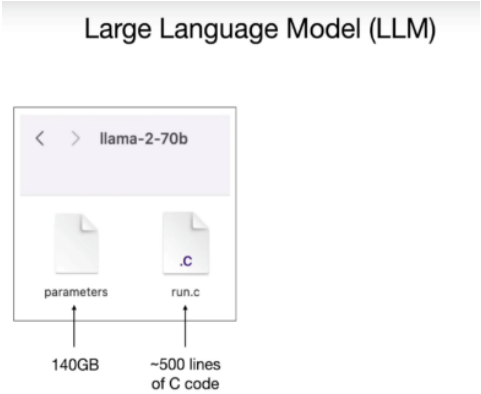
数据来源：微软 Build 大会

图 12 卷积运算的过程



数据来源：cnblogs

图 13 LLM 的参数文件和运行代码



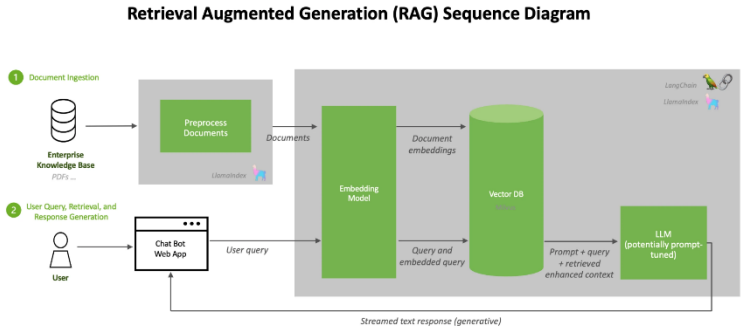
数据来源：华尔街见闻

大模型依靠数据和算力将人工智能提升到前所未有的高度，但依然存在存在 1）输出可靠性不稳定 2) 结果难以解释，3）数理能力较弱等弊端。由于大模型的参数规模较大，不仅抬升了存储成本，而且需要使用更强的算力资源，极大抬升了大模型研发的资本支出。即便如此，从另一个维度看，大模型的处理能力主要依赖于已有的训练数据，并且若数据本身存在缺陷或是错误，亦会影响大模型的训练效果，这也将直接影响模型对于上下文的理解和对话的一致性。大模型的数学和逻辑推理能力仍然需要加强。诸多大模型在执行推理任务的过程中并没有提供其信息的来源，因此验证输出可靠性具备挑战。

虽然依托人工智能的大模型技术还不尽完善，但是展现出了巨大的潜力，技术发展上呈现了很多共同的趋势如：多模态，轻量化，自增强，增加逻辑推理及互联和 API 化。

我们认为 RAG 检索增强生成技术，使得不那么完美的基于 LLM 的 AI 成为一种确信可用的技术，基于准确性和可追溯性的考量，AI 中的检索增强生成（RAG）是一种变革性范式，通过使用从外部来源获取的事实，以此提高生成式 AI 模型的准确性和可靠性。据英伟达官网，RAG 架构的工作流程可概括为：当用户向 LLM 提问时，AI 模型会将查询发送给另一个模型，后者会将查询转换成数字格式以便机器读取；随后嵌入模型将这些数字与可用知识库的机器可读索引中的向量进行比较，检索相关数据并发送回 LLM；LLM 将检索到的单词和它自己对查询的响应相结合，形成最终的答案并提交给用户，其中可能会引用嵌入模型找到的来源。

图 14 RAG 架构的工作流程



数据来源：thenewstack

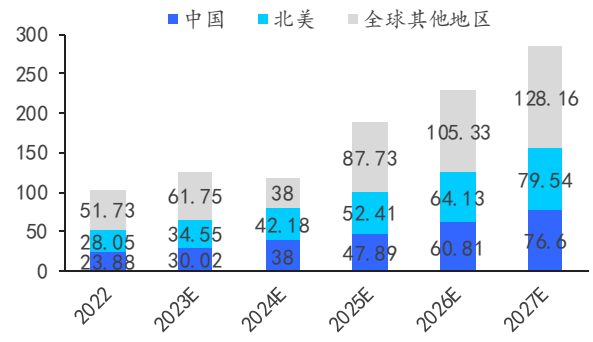
2.2. 通用计算转向加速计算

2.2.1. 数据特征与 AI 计算

数据规模的激增和数据类型的转变需要更高的计算效能。据 IDC 预测，全球数据规模 2027 年或将高达约 284ZB（十万亿亿字节，泽字节）。而随着移动互联网的普及，叠加物联网、智能手机和可穿戴设备的发展，全球设备和人群社交等的广泛“联网”驱动了个体间的数据互动，并产生巨量的非结构化数据，而其中多媒体数据的占比显著提高。据 Statista，2022 年 6 月，谷歌公司旗下的全世界最大视频网站 Youtube 每分钟全球上传的视频时长高达 500 小时。视频数据中信息的识别难度远高于文本数据和图像数据。在非结构化数据大量涌现的当下，更高的计算效率成

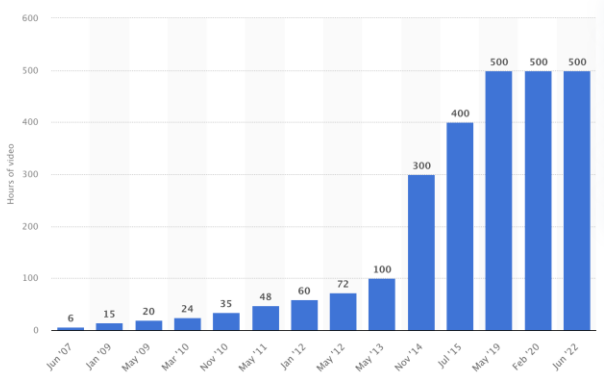
为迫切需求。

图 15 全球数据量激增（单位：ZB）



数据来源：IDC，国泰君安证券研究

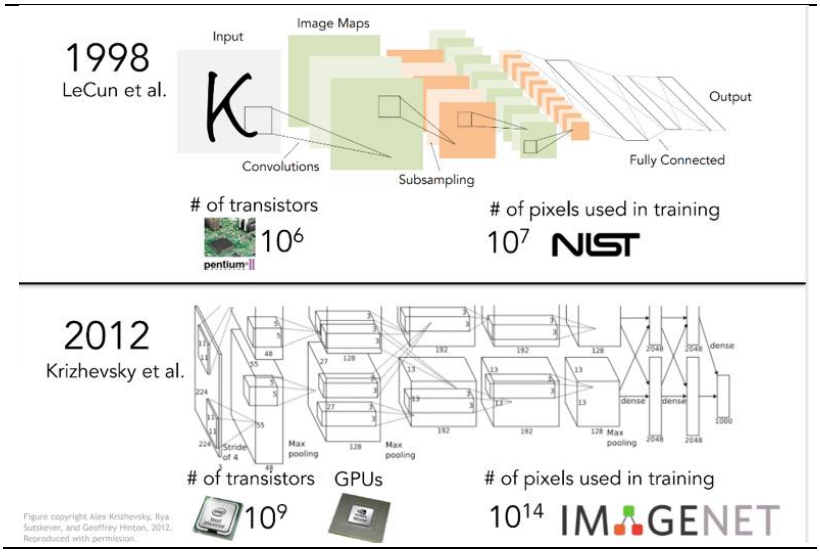
图 16 YouTube 每分钟上传的视频时长



数据来源：Statista

复杂卷积神经网络需要高效能算力实现。1998 年，LeNet-5 推出，主要用于解决手写的图像识别问题，奠定了现代卷积神经网络的基本结构：卷积、池化、全链接。从现实需求看，近年来图像数据的数量和图像的像素点大幅增多，图像识别的需求进一步提升；LeNet-5 在 1998 年使用 Intel 奔腾 II 处理器训练，将文字识别率提升到 99%，但无法识别图像；在 2012 年 AlexNet 延续了 LeNet-5 的架构，修改了激活函数,增加了训练数据集,使得 ImageNet 的的识别率比 2011 年陡然提升,提升到 84%，训练使用 CPU 晶体管数量是原 1998 年的 1000 倍，外加两块 NVDA 的 GTX580，每颗有 30 亿颗晶体管的并行算力。

图 17　复杂卷积神经网络需要高效能算力实现

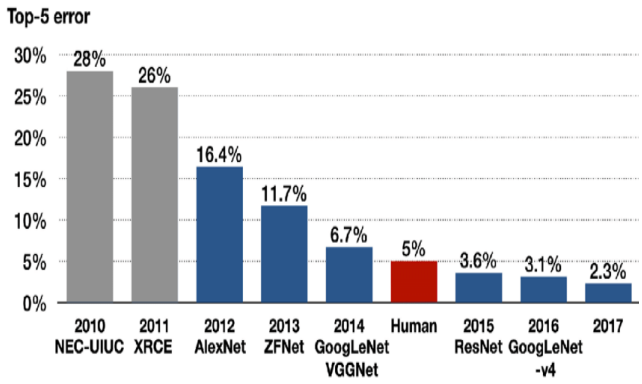


数据来源：51CTO

更多的训练参数、更大的算力会驱动神经网络的识别度提高。我们以赢得 2010-2017 年 ImageNet 大规模视觉识别挑战赛 (ILSVRC) 的算法为例，其中 Top-5 错误是指算法对图像提出的所有 top-5 分类都是错误的概率，蓝色柱状图的算法均是卷积神经网络。其中，人类的识别错误率约 5%，亦即从 2015 年以来基于卷积神经网络的算法已超过人类平均的视觉识别能力。这正是由于更多的训练参数、更大的算力会驱动神经网

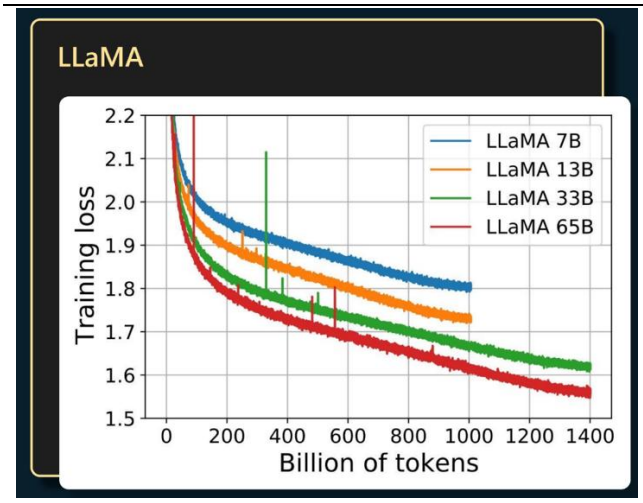
络的识别度提高。在大模型的实际预训练中，研究者监测损失函数确定模型迭代中的表现变化，损失低说明 Transformer 更可能给出正确预测，随着大模型参数的递增，大模型的表现普遍愈好，这正是算力提升的结果。

图 18 先进大模型的识别能力已高于人类



数据来源：ResearchGate

图 19 LLAMA 的训练损失



数据来源：微软 Build 大会，量子位

2.2.2. 加速计算

CPU 与 GPU 的组成结构与运算方式存在不同。CPU 作为电脑的中央处理器，其运行主要以串行计算的方式进行。串行计算指的是多个程序在同一个处理器上被执行，只有在当前的程序执行结束后，下一个程序才能开始执行。而 GPU 是电脑的图形处理器，最初主要用于图像运算，适用于加速计算场景中实现并行计算的需要。二者组成的不同主要表现在，CPU 拥有更大的逻辑运算单元和控制单元，同时拥有更大的缓存空间，但 GPU 却拥有更多的逻辑运算单元数量。

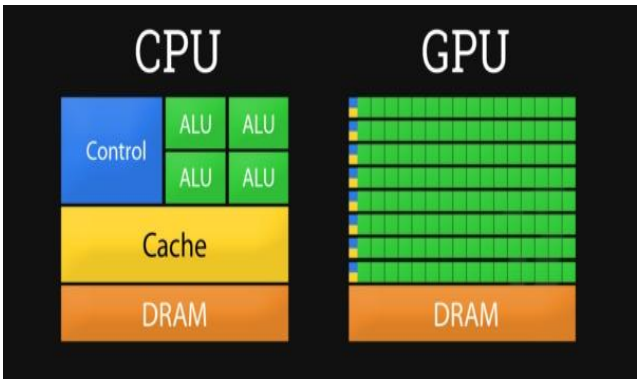
表 1 GPU 和 CPU 的区别

对比项	CPU	GPU
设计理念	CPU 设计注重通用性和灵活性，适合处理复杂的、串行的计算任务	GPU 设计重点在于处理大量的并行任务，适合执行重复且简单的操作
核心结构	CPU 通常包含较少的核心，但每个核心能够处理复杂任务和多任务并发	GPU 包含成百上千的小核心，每个核心专注于执行单一任务，但在并行处理大量数据时表现卓越
处理速度	在执行逻辑复杂、依赖于单线程性能的任务时，CPU 通常表现更优	GPU 在处理可以并行化的大规模数据时，如图像处理、科学计算，表现出远超 CPU 的处理速度
能效比	在单线程任务中，CPU 提供更高的能效比	当任务可以并行化时，GPU 在能效比上通常更有优势，尤其是在大规模计算任务中
优势场景	**复杂逻辑处理**：适合处理需要复杂决策树和分支预测的任务，如数据库查询、服务器应用等 **单线程性能要求高的任务**：在需要强大单线程性能的应用中，如某些类型的游戏或应用程序	**数据并行处理：**需要同时处理大量数据的场景下，如深度学习、大规模图像或视频处理 **高吞吐量计算任务：**适用于需要高吞吐量计算的应用，如科学模拟、天气预测等

数据来源：techlead，国泰君安证券研究

大模型算法可分拆成并行计算，GPU 可适应神经网络训练高并发、并行计算和矩阵处理需要。LLM 大多基于神经网络，大模型军备竞赛导致开发者对 LLM 推理性能的要求激增，因此神经网络参数数量迅速飙升。由于大模型基于的神经网络是高度并行的，使用神经网络做的许多计算都可以分解成更小的计算单元，以此执行串行计算。CPU 内部运算单元有限，在执行矩阵运算时将极大的消耗模型训练的时间，因此使用 GPU 进行大规模并行计算的优势得到了充分彰显，以 H100 Tensor Core GPU 为例，其支持多达 18 个 NVLink 连接，总吞吐量为 900 GB/s，是 PCIe 5.0 带宽的 7 倍，进而实现超快速的深度学习训练。GPU 逻辑运算单元较多的优势在加速计算等场景中能够得到充分的发挥，因此 GPU 无疑成为训练 LLM 的首位硬件选择。

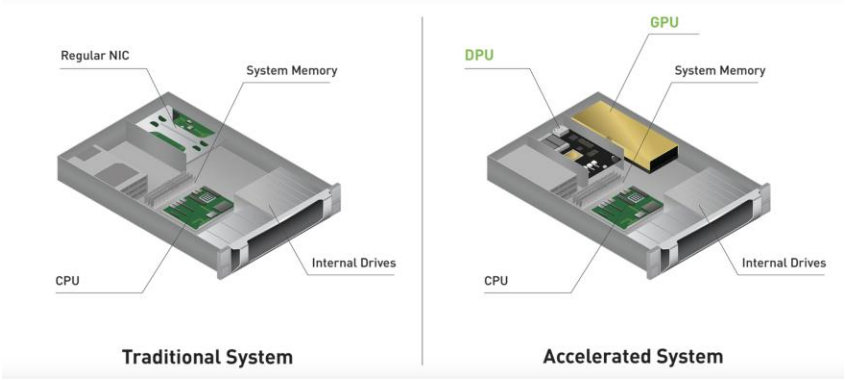
图 20 CPU 与 GPU 的结构区别



数据来源：腾讯云开发者社区

通用计算向加速计算的转型是迈向 AGI 的必由之路。加速计算作为现代计算方式，能够将应用的数据密集型部分分离，在一个单独的加速器件上处理。硬件端，加速计算利用 GPU、DPU、TPU、ASIC 和 FPGA 等性能、结构各异的计算单元来执行比 CPU 更高效的计算。加速计算解决方案需要硬件、软件和网络三者的协同，每个加速平台硬件系统都根据不同用例需求而设计，并辅之以针对具体业务应用的运营和管理的软件栈。作为计算转型的下一个阶段，就像如今的智能手机都逐步应用混合 AI 技术，开始陆续配备能在终端运行大模型的高性能 GPU 一样，未来每个服务器和工作站都将配备计算加速器，为更多的计算密集型工作提供支持，服务于包括企业、主机托管、云、边缘和模块化设施等在内的数据中心。

图 21 加速计算为 AI 的现代应用提供支持



数据来源：英伟达官网

3. 不仅是芯片，而是计算平台型公司

3.1. GPU 引领加速计算

AI 迭代飞速催生芯片技术创新，TPU、FPGA、ASIC 等 AI 芯片应时代需求而生。AI 迭代飞速催生芯片技术创新，FPGA 和 ASIC 等专用芯片市场份额有所提升，例如谷歌推出的 TPU (Tensor Processing Unit)，与 Tensorflow 框架紧密集成，专为在大规模分布式系统中进行深度学习训练和推理。但 FPGA 和 ASIC 的通用性较差，通常针对特定任务效率较高。目前 GPU 已在大模型的训练和推理侧均形成显著的竞争优势，且我们认为，基于通用计算向加速计算转型的趋势，GPU 具备更强的综合性能和通用性。

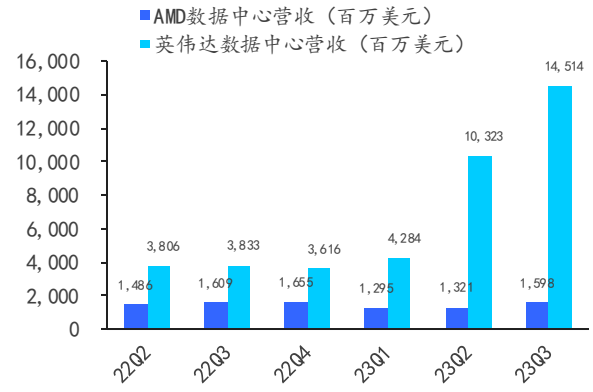
表 2 TPU、FPGA 和 ASIC 等 AI 芯片技术爆发式增长

芯片	中文名称	训练效率	训练速度	推理效率	推理速度	通用性	推理准确性
CPU	中央处理器	1 倍基准	1 倍基准	1 倍基准	1 倍基准	非常高	98%-99.7%
GPU	图像处理器	10-100 倍	10 到 1000 倍	1-10 倍	1-100 倍	高	98%-99.7%
FPGA	现场可编程门阵列	-	-	10-100 倍	10-100 倍	中等	95%-99%
ASIC	专用集成电路	100-1000 倍	10-1000 倍	100-1000 倍	10-1000 倍	低	90%-98%

数据来源：AILI，国泰君安证券研究

GPU 引领加速计算转型，驱动英伟达营收高增，显著高于 AMD。英伟达以游戏显卡业务起家，随着生成式 AI 和向加速计算转型的催化，数据中心逐渐变成主营业务。FY24Q3（对应自然年约 23Q3）英伟达营业收入 181.20 亿美元，其中数据中心收入达 145.14 亿美元，同增 278.7%，占总营收的 80.1%。对比主要竞对 AMD，备受市场关注的数据中心 GPU MI 系列或将在 2024 年尚可实现全面出货，因此其 23Q3 数据中心收入仅 15.98 亿美元，其中绝大部分为 X86 服务器收入，同增 21.0%，占总营收 27.6%。

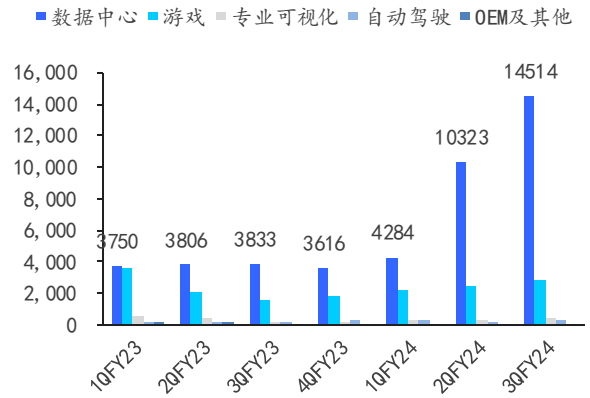
图 22 英伟达数据中心营收大幅领先 AMD



数据来源：公司财报，国泰君安证券研究

注：英伟达财务数据经近似调整为自然年

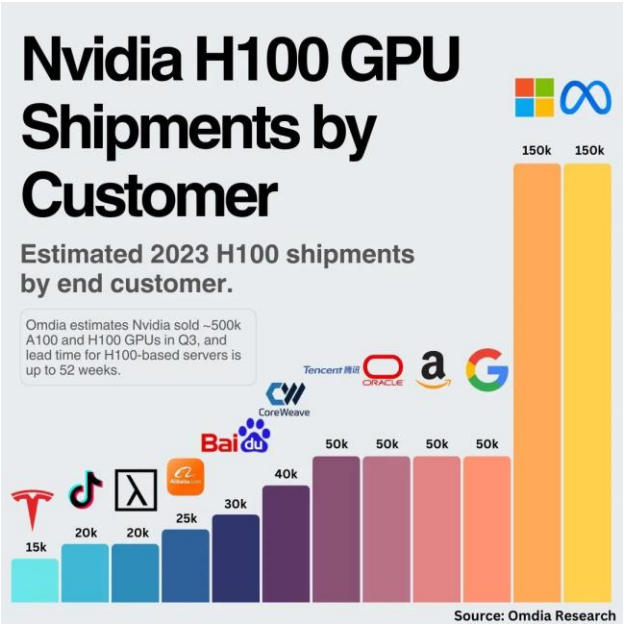
图 23 数据中心驱动英伟达营收高增（单位：百万美元）



数据来源：公司财报，国泰君安证券研究

AI 和加速计算爆火的趋势下，GPU 的争夺格外激烈。据市场跟踪公司 Omdia 的统计，英伟达在 23Q3 大约卖出了 50 万片 A100 和 H100 GPU。从全球头部云服务商的采购情况看，Meta 和微软是最大买家，分别采购了多达约 15 万个 H100 GPU，大大超过了谷歌、亚马逊、甲骨文和腾讯约五万个的采购数量。庞大的需求量也使得基于 H100 需约 36-52 周的时间才能交付。

图 24 23Q3 英伟达 A100&H100 出货量约 50 万片



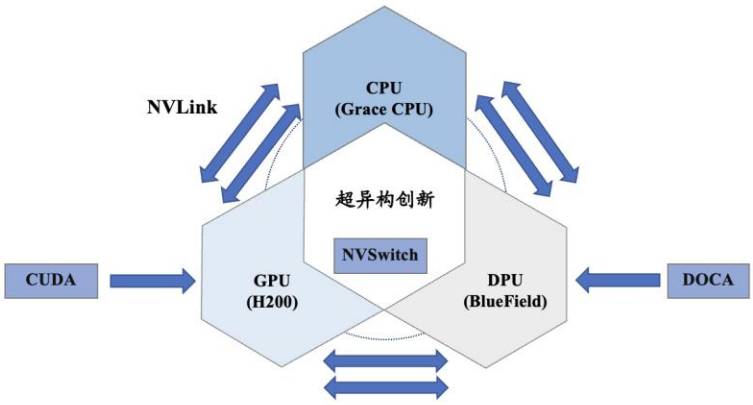
数据来源：Omdia Research，新智元

3.2. 超异构计算演进系统优势

英伟达以超异构创新构建面向大规模 AI 计算的超级计算机。英伟达构建了面向大规模并行计算而设的全栈异构的数据中心解决方案。两种不同的芯片在一起工作叫异构，将三种不同的芯片通过特殊通信连接成系统一起工作叫超异构，随着 GPU 并行计算的加速计算处理，数据的计算

结果使得的 GPU 每秒的数据吞吐量剧增，原来基于传统 X86 服务器的 PCIe 通信能力制约了整个计算系统的速率，所以英伟达以加速计算为中心，重构新型计算系统，自己设计了基于 ARM 架构的 Grace GPU,Bluefied DPU(提高内存访问速度，协议及安全解析效率)，并设计了 NVlink 通信系统，将这三种芯片的片间通信提升到能力提升到 900G/s，超过原来 PCIe 通信能力的 7 倍。

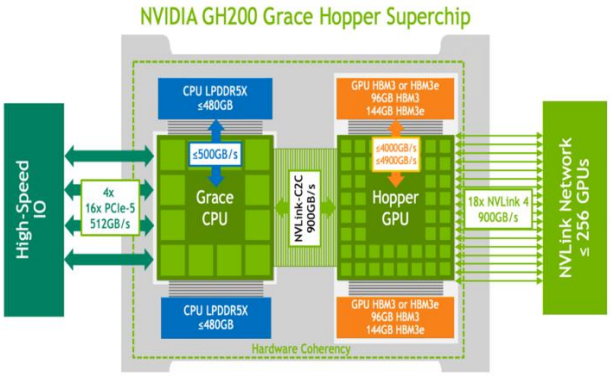
图 25 英伟达超异构创新整体框架



数据来源：英伟达官网，国泰君安证券研究

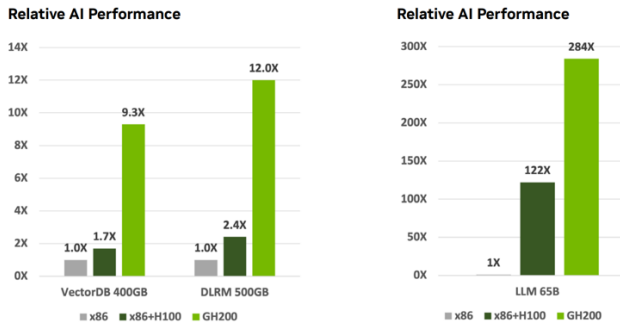
GH200 是真正适用于 HPC 工作负载的异构加速平台。以 GH200 超级芯片为例，其使用 900GB/S 的 NVLink-C2C 芯片互连，将基于 Arm 的 Grace CPU 与 H100 Tensor Core GPU 整合，从而不再需要传统的 CPU 至 GPU PCIe 连接。CPU 和 GPU 分别配备 480GB 的 LPDDR5X 内存和 96GB 的 HBM3 或 144GB 的 HBM3e 内存，集成高达至少 576GB 的高速访问内存，可流畅运行具有数 TB 大小的嵌入表的推荐系统和向量数据库。从官网公布的性能上看，其在 VectorDB 向量数据库、DLRM 推荐算法模型和 LLM 上的表现均大幅高于 x86 及 x86+H100。我们认为，GH200 专为加速计算工作负载而设计，是适用于 HPC 工作负载的异构加速平台。

图 26 英伟达 GH200 超级芯片



数据来源：英伟达官网

图 27 GH200 相对 AI 性能显著高于 x86 及 H100

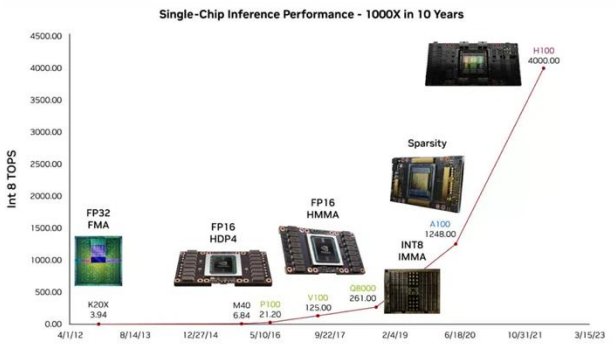


数据来源：英伟达官网

摩尔定律后期，黄氏定律重塑半导体行业准则，英伟达芯片的迭代是与

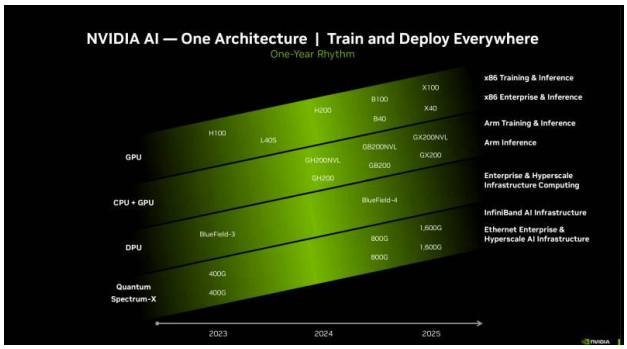
自己赛跑的过程。摩尔定律指在价格不变的前提下，集成电路上可容纳的晶体管的数目，每隔约 18 个月便会增加一倍。但随着传统半导体晶体管结构已进入纳米级别，叠加大模型对于算力激增的需求已远大于摩尔定律所预估。黄仁勋对 AI 性能的提升作出预测，指出 GPU 将推动 AI 性能实现每 1 年翻 1 倍，也就是每 10 年 GPU 性能将增长超 1000 倍。这一论断也被称之为“黄氏定律”。复盘过去十余年间的英伟达芯片性能，H100 已经将推理性能拓展至 1000 倍，而根据英伟达的战略规划，在未来两年，无论是 GPU、CPU 还是 DPU，也均会有下一代产品发布。

图 28 黄氏定律



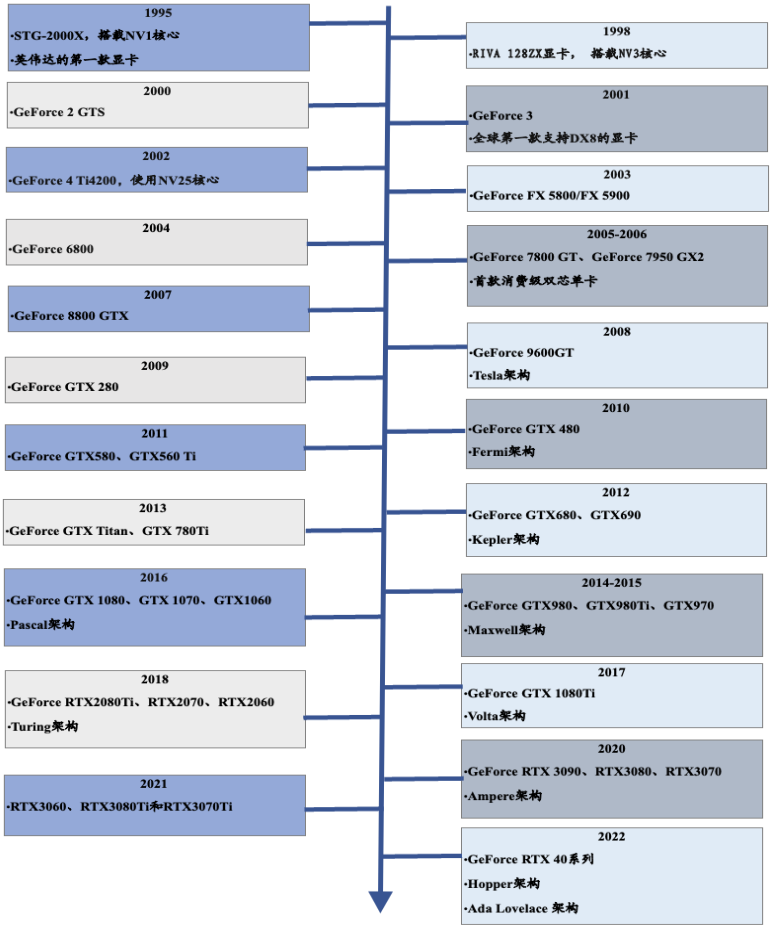
数据来源：英伟达官网

图 29 未来英伟达芯片产品仍将持续迭代



数据来源：英伟达

图 30 十余年间英伟达芯片架构飞速迭代



数据来源：英伟达官网，CSDN，国泰君安证券研究

3.3. 不仅是芯片，而是计算平台型公司

通过分析 AI 系统全栈架构，我们认为英伟达无疑已经成为一家提供超级计算平台的公司。我们认为，英伟达的增长优势并不仅像市场认为的基于大模型训练和推理 GPU 算力需求，要从计算平台型公司的视角充分论证英伟达的核心竞争力。AI 系统由下至上分别由体系结构、编译器、框架、开发和应用构成。英伟达的产品和服务体系几乎贯穿了整个 AI 系统全栈架构，从 GPU、超级计算机和网络加速器，到 CUDA 编译器，到 AI 推理引擎等。即便在应用侧，英伟达也发布了自己的大语言模型 ChipNeMo，据英伟达官方博客，ChipNeMo 可提供聊天机器人、代码生成器和分析工具，帮助提高芯片设计效率。

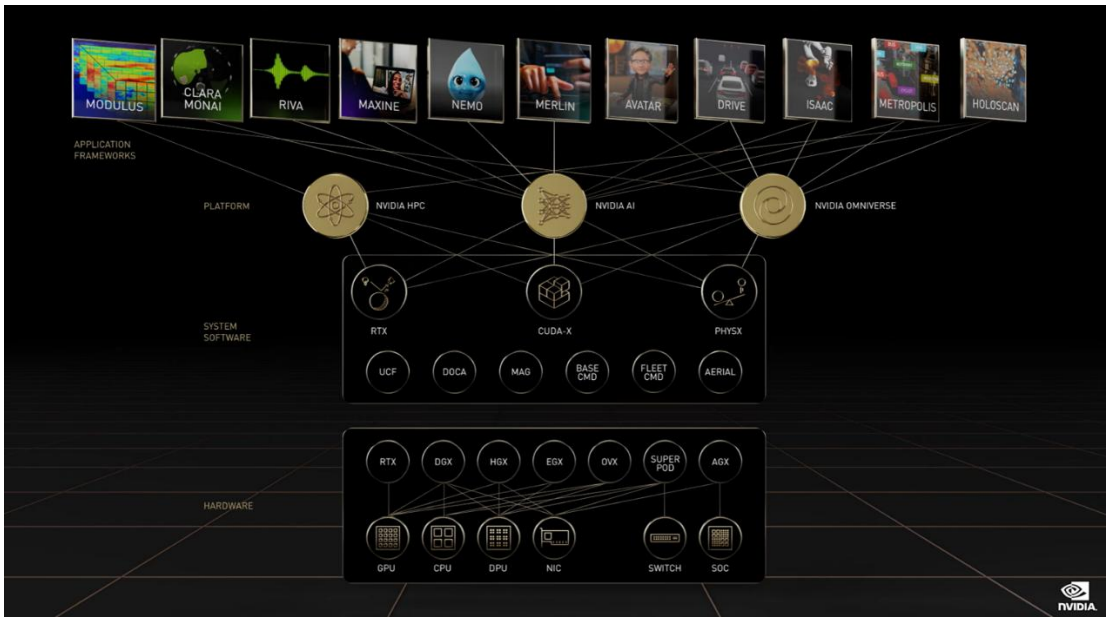
图 31 AI 系统全栈架构图

数据来源：Github

3.3.1. 基于产品的平台

英伟达全面的产品矩阵优势为计算平台的核心竞争力发挥提供了保证。英伟达通过平台型产品结构设计打造规模效应。详细拆分，我们可将英伟达划分为硬件、软件、平台、应用框架四个维度。英伟达基于“硬件+软件”的技术优势，同时依托面向行业打造的应用框架，提供了对于细分行业定制的行业解决方案。

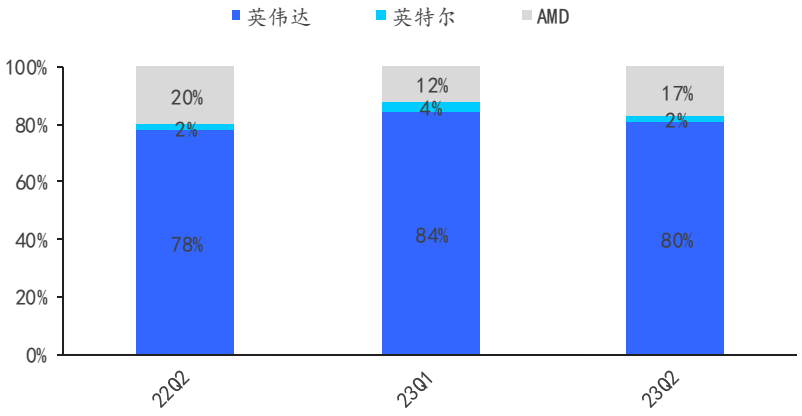
图 32 英伟达产品矩阵



数据来源：英伟达 GTC 大会

英伟达独显市场份额长期稳居高位，与 AMD 此消彼长。据 JPR，23Q2 全球独立显卡出货量为 640 万块。其中 23Q2 英伟达市场份额高达约 80%，AMD 仅约 17%。

图 33 23Q2 独立显卡市场份额



数据来源：JPR，国泰君安证券研究

3.3.2. 基于技术的平台

除了硬件上的系统性优势，软件端英伟达亦通过 CUDA 打造高兼容性的 GPU 通用平台，依托系统和生态形成竞争优势。以 CUDA 为例，CUDA 是英伟达开发的并行计算平台和 API 模型，是加速计算密集型任务最为依赖的软件生态基石。以 AMD 的 ROCm 作为对比，CUDA 具备 ROCm 难以企及的开发人员数量，目前 CUDA 拥有超过 400 万开发人员，历史上 CUDA 总下载量达到 4000 万，CUDA 已形成社区支持和成熟的生态体系，使得 GPU 的可编程性得到了飞跃。整体而言，在发布时间、硬件支持、操作系统和开发者数量等维度上 CUDA 均具备优势，展现出更加繁荣的生态。此外，Infiniband 提供网络解决方案，通过提供网卡、DPU、交换机等产品以适应处理高分辨率模拟、超大型数据集和高

度并行的算法的需要，据 FY24Q3 业绩会，英伟达网络业务已经达到 100 多亿美元的规模。

表 3 CUDA 较 ROCm 仍具备明显生态优势

对比项	CUDA	ROCm
发布时间	2007 年	2016 年
硬件支持	支持 2006 年后的所有英伟达 GPU	仅支持 AMD 高端 GPU 系列，直至 223 年 4 月开始逐步向消费级 GPU 拓展
操作系统	Linux 和 Windows	原本仅支持 Linux，2023 年 4 月开始逐步支持 Windows
开发者数量	超过 400 万开发人员，Github 上有超 35100 个开发者发布的软件包库	使用人数较少，Github 上仅有 600 余个开发者发布的软件包库

数据来源：英伟达，AMD，Github，国泰君安证券研究

3.3.3. 服务客户的平台

作为平台型公司强调技术可拓展性，通过广大的客户群体和产业链合作伙伴加速全球 AI 生态建立，以适应由加速计算带来科技变革。1）从客户数角度，黄仁勋表示已有 15,000 家初创公司建立在英伟达的平台上，全球有 40,000 家大型企业正在使用加速计算，FY24Q2 电话会中，黄仁勋表示未来四年将有约 1 万亿美元用于 AI 数据中心升级，每年的资本支出约为 2500 亿美元，其中包括 GPU 的升级；2）从产业链合作伙伴角度，英伟达作为 IC 设计领军者，与下游晶圆代工、封装测试、显卡组装等产商建立了稳定的合作关系，积极响应数据中心、游戏、汽车和专业可视化等各类终端客户需求，在近 30 年来通过合作伙伴网络（NPN）将产品投入市场；3）从投资角度，以英伟达企业投资部门为例，2023 年来已投资芯片到芯片光纤连接的 Ayar Labs、AI 模型中心 Hugging Face、云服务提供商 CoreWeave 在内的 14 个项目。

图 34 英伟达打造“服务客户的平台”

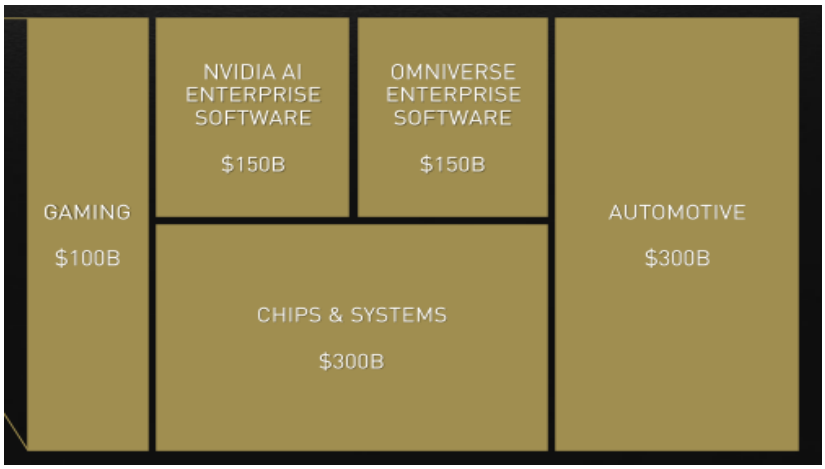
数据来源：国泰君安证券研究

4. 计算转型初期，强化超线性增长

4.1. 加速计算市场规模测算

算力板块正在同时经历两个临界点——加速计算和生成式 AI，企业竞相将生成式 AI 应用到各个产品、服务和业务流程中，AI 加速渗透。由于在 AI 大模型的竞赛中，英伟达目前处于供应算力的绝对领导优势，我们持续看好 GPU 未来的市场空间。我们从几个维度做了一个参考，几个数据维度做了一个相应的测算。同时，从 24 年前三个季度表现及市场的对 24 财年的一致预期推测，我们认为 2023 年英伟达的数据中心收入约为 463 亿美元，除去网络设备的 100 亿美元，数据中心的 GPU 约为 363 亿美元，根据英伟达预计其综合业务领域的总潜在市场(TAM) 为 1 万亿美元，分布于游戏、汽车、芯片和软件,其中芯片和系统达到 3000 亿美金，AMD 的 CEO 苏姿丰预测数据中心 AI 芯片的 2023 年总市值将突破 450 亿美元，远高于 AMD 6 月预测的 300 亿美元。到 2027 年，以 70%的 CAGR,数据中心加速计算市场将进一步扩张至 4000 亿美元。站在 2023 年，英伟达剔除网络设备外，GPU 收入在 360 亿美金左右，根据英伟达和 AMD 的预测，我们假设 2027 年 NVDA 占据数据中心 GPU 市场 80%，NVDA 的 GPU 将可能以 70%的 CAGR 在 2027 年实现 3000 亿美金收入,2023 年除 GPU 外的数据中心网络设备有 100 亿美金的收入，综合考虑到这些维度的收入，我们预计英伟达数据中心业务整体在 2027 年有望实现 3000 亿美金的收入。

图 35 NVDA 总潜在市场（TAM）为 1 万亿美金

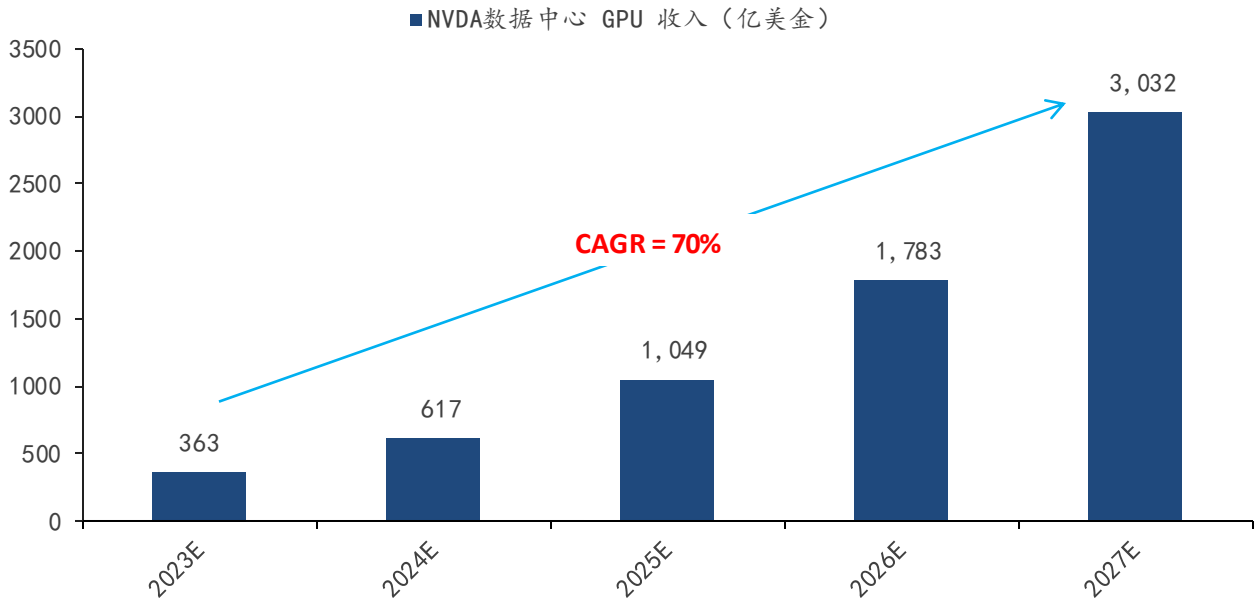


数据来源：英伟达官网

图 36 苏姿丰预测数据中心 AI 芯片 2027 年总市值将超 4000 亿美元

数据来源：至顶网

图 37 2023 年-2027 年英伟达数据中心 GPU 将保持高增，CAGR 达 70%



数据来源：英伟达，AMD，国泰君安证券研究

4.2. 技术普及推动全面增长

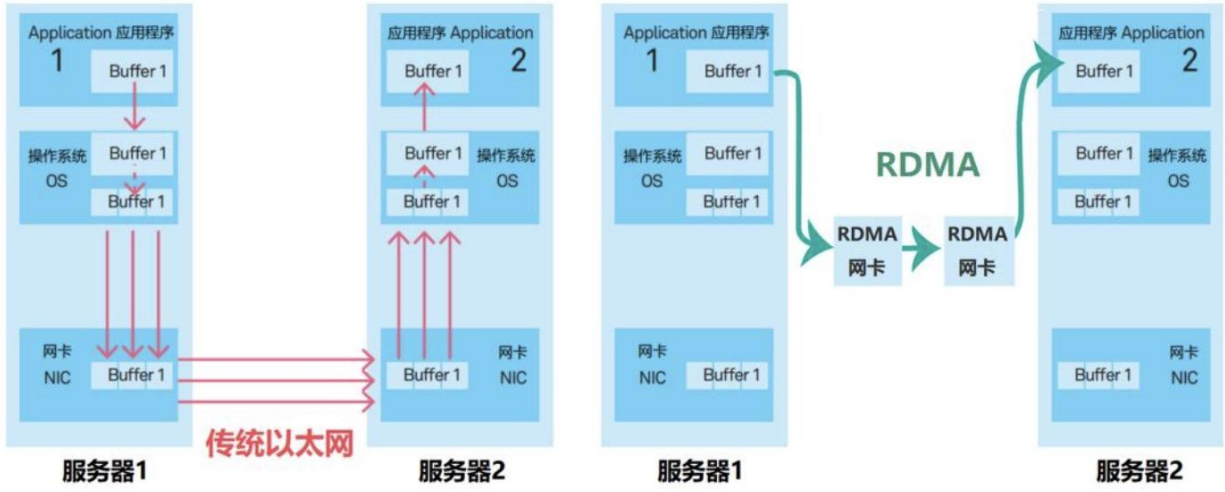
4.2.1. 加速计算技术生态系统迭代增长

GPU 的高增往往忽视了以 INFINI BAND 领衔的全栈周边网络设备的增长。随着人工智能和生成内容的大型模型的兴起，对高性能计算和智能计算的需求急剧增加。为了满足这些巨大的计算需求，建立高性能计算集群成为了必然选择。在这种背景下，InfiniBand 因其卓越的性能表现，成为搭建高性能计算集群的首选技术。

InfiniBand 作为一种高性能的计算机网络通信标准，主要用于高速数据传输。它在高性能计算 (HPC)中特别受欢迎：1）**高带宽和低延迟**：InfiniBand 提供了非常高的数据传输速率，远高于传统的以太网。同时，它的延迟极低，这对于要求实时数据处理的 AI 应用至关重要。2）**可扩展性**：InfiniBand 支持构建大规模网络，非常适合目前英伟达的大型计算集群和数据中心。3）**灵活性和多用途性**：InfiniBand 支持 RDMA，允许网络设备直接访问应用程序的内存，无需 CPU 介入，从而减少延迟并提高吞吐量。它不仅用于高性能计算，还可以在存储网络和数据中心网络中发挥重要作用。InfiniBand 的这些特性使它成为了解决高性能计算和数据密集型中的数据传输挑战的理想选择。

我们看好受到数据中心和人工智能的强劲需求推动下， InfiniBand 的未来的潜力充满期待。根据 36Kr 的报道，到 2029 年，InfiniBand 的市场规模将达到 983.7 亿美元，相比 2021 年的 66.6 亿美元，增长 14.7 倍。预测期内（2021-2029）的复合年增长率，为 40%。

图 38 RDMA 相当于是一个“消灭中间商”的技术



传统模式 VS RDMA模式

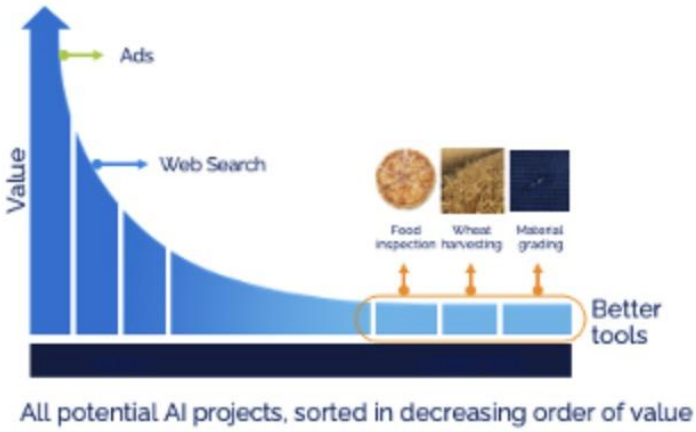
数据来源：36Kr

除核心 GPU 芯片组件和 INFINIT BAND 带来的增长以外，英伟达的爆发式增长还带来周边网络设备的增长如，SPACTRUM ，ENTERPRISE AI ,AI TOOLS chain。Nvidia 推出的 Spectrum-X 技术，专为 AI 工作负载设计，提供高于传统以太网 1.6 倍的性能。该技术整合了 Nvidia 的 H100 Tensor Core GPU、AI Enterprise 和 AI Workbench 软件，与戴尔、HPE 和联想的新服务器产品结合，旨在为企业提供执行高级生成式 AI 模型的完整解决方案。Spectrum-X 结合了 Spectrum-4 以太网交换机和 BlueField-3 SuperNIC，一个新型网络加速器，以提高 AI 工作负载处理速度和安全性。这些新系统预计 2024 年第一季度上市，标志着向生成式 AI 时代的重要转变。Nvidia 在其 Israel-1 超级计算机上已部署 Spectrum-X，为构建下一代 AI 系统的公司提供参考。戴尔、HPE 和联想的合作强调了为客户提供高性能 AI 解决方案的重要性。

4.2.2. 应用、行业和区域的技术扩散增长

生成式 AI 落地，我们认为将从大规模用户向小众定制化行业进行传导。最先受益的，就是各大企业的广告业务，这些企业希望推出各种工具来帮助企业降低成本。对数字营销较为倚重的行业也因此被改变。在 AI 的帮助下，不管是文字宣传还是图片，甚至是视频制作，效率都大大提高，制作一条产品宣传视频时间仅仅需要 6-8 分钟，目前一些从业者的成本仅仅来自素材和推广，大大地节省了制作的成本。随即搜索也慢慢的被大模型渗透，生成式 AI 搜索直击用户痛点。根据用户输入的关键词或问题，以答案的形式为用户提供一个文本结果，节省时间和精力。随着不同行业有着各自独特的需求和挑战。生成式 AI 开始被定制化，以适应特定行业的需求。在竞争日益激烈的市场环境中，行业例如医疗、法律、教育和金融服务等寻求通过定制化 AI 应用来获得竞争优势，也推动了技术创新。

图 39 生成式 AI 从大规模用户向小众定制化行业进行传导



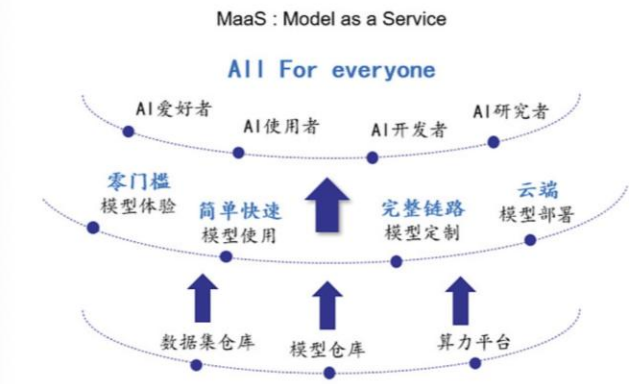
数据来源：Andrew Ng

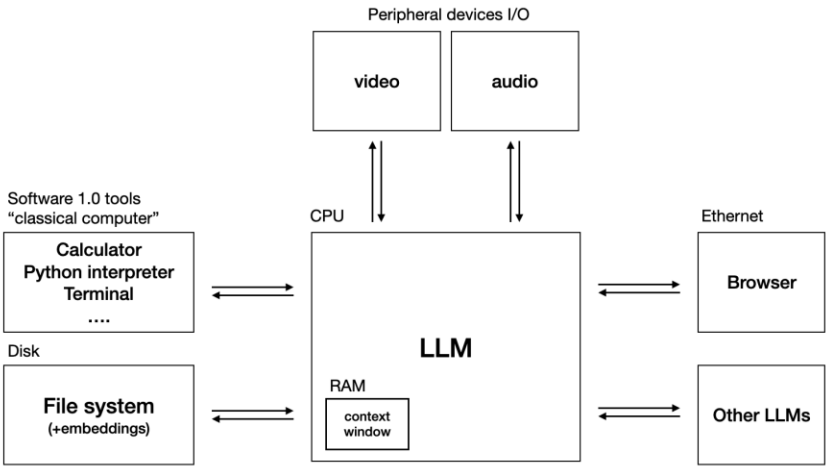
Maas 催化企业能够方便地使用高级的 AI 模型，提升渗透率。
中小企业计划引入人工智能来提高效率，但面临几个主要问题：

1）**高开发门槛：** AI 模型开发需要大量数据和强大的 GPU 算力，通常只有大公司才能承担这样的投入，对中小企业来说是一个挑战。
2）**定制化需求：** 虽然有通用的模型，但在具体问题上，效果可能不佳，需要进行定制化调整，这对开发人员提出了较高的要求。
3）**模型多样性和调用复杂性：** 现有的众多模型各有不同的调用方式，开发者需要查阅大量资料并调整各种参数，这增加了使用不同模型的难度。

Model as a Service (MaaS) 成功解决了这些问题。MaaS 是一种服务模式，其中云计算提供商（如阿里云、腾讯云等）不仅提供硬件资源、通用软件能力和底层框架，而且还将模型作为服务的一部分。这样，模型成为了一种重要资源，可以更容易地被企业利用，而不需要自己从头开始开发。MaaS 的目标是实现“All For Everyone”，即让所有企业都能够方便地使用高级的 AI 模型。

图 40 Maas：Model as a Service 降低公司 AI 使用门槛





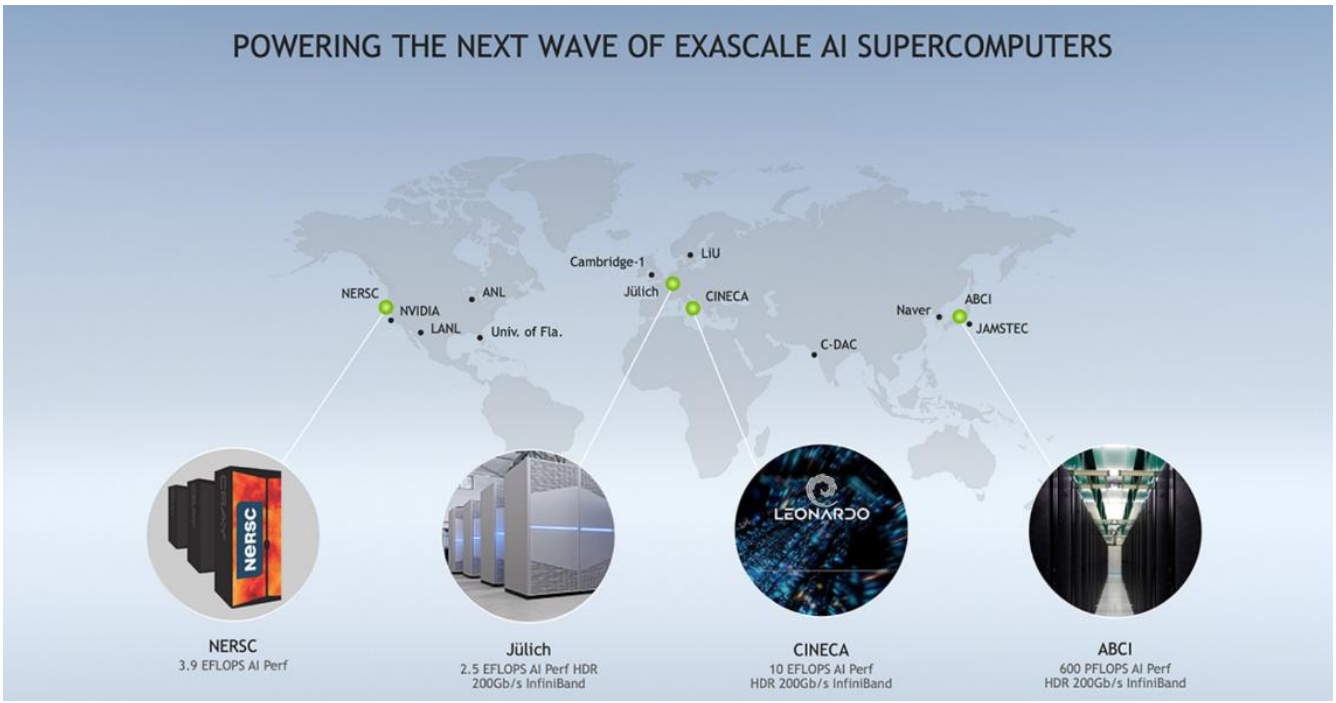
数据来源：阿里云栖大会，国泰君安证券研究

随着大型语言模型引发的革命性变化，操作系统的领域有了重塑的可能。操作系统不仅是释放硬件性能的关键入口，同时也是众多软件服务的基础平台，它构成了所有人机交互的初始点。结合了传统操作系统功能和针对 AI 应用的特定需求的 AI 操作系统，为不同规模和类型的企业提供了一个功能强大、高效且用户友好的平台，使他们能够最大限度地发挥 AI 技术的潜能。

而基于 AI 操作系统的 AI 应用是更大的机会，也将带来成倍的算力需求。AI 软件应用的蓬勃发展可以形容为一场技术革命，它正在迅速渗透到社会的各个角落。AI 技术的通用性使其能够被应用于多个行业，从医疗健康、金融服务到教育、交通等，这种跨行业的适用性极大地扩展了 AI 软件应用的范围。

由于算力在国家经济中的拉动作用具有重要意义，各国之间形成了在这一领域的竞争。由于计算能力在全球经济中扮演着至关重要的角色，各国都在积极参与这场科技竞赛。例如，中美之间在人工智能领域的竞争尤为激烈。美国为了限制中国人工智能的发展，对中国的高端计算芯片实施了出口限制。同样，世界上其他一些国家也加入了这场竞赛，例如英国购买了大量英伟达的 GPU，并计划建立运算中心，德国和法国等国也在进行类似的努力。目前，英伟达已经支持了全球多个超级计算中心的建设和运营。。

图 41 各个国家在算力的竞争趋于白热化

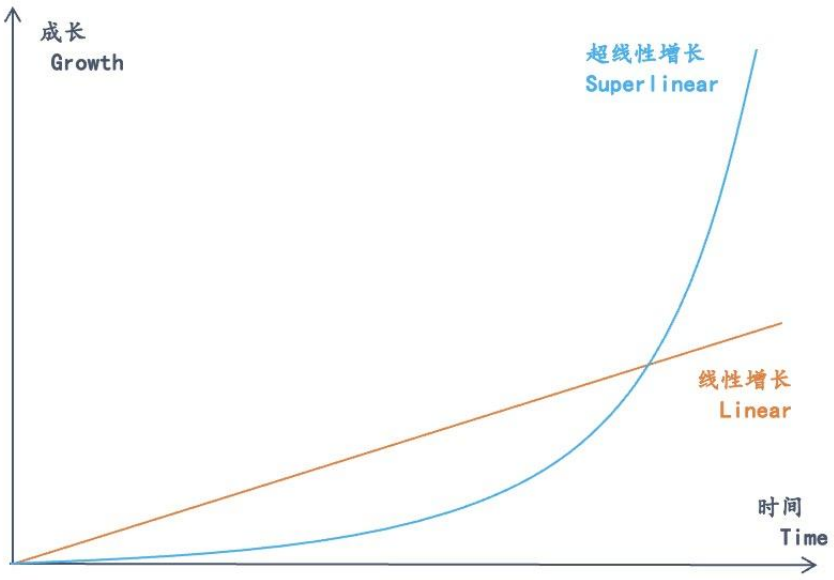


数据来源：英伟达官网

4.3. 超线性增长

超线性（Superlinear）的增长，用于描述量与量之间的一种变化关系，例如 y=a+b*x^n，其中 n＞1。当 n=1 时，表示为线性关系，当 0＜n＜1 时，表示为亚线性关系。超线性增长在自然界中普遍存在，尤其适用于平台型互联网公司及技术的扩散初期，特征表现在：

图 42 超线性增长与线性增长

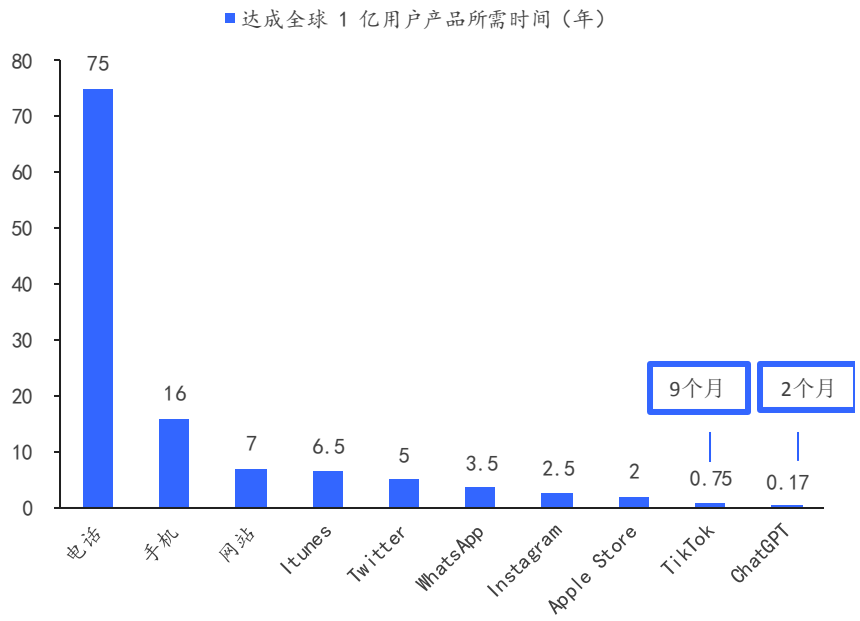


数据来源：国泰君安证券研究

1）单一维度超过线性增长，超过一定阈值，增长会随着变量而加速，我们看到上图，就人类技术整体而言是加速普及的，速度越来越快的超线

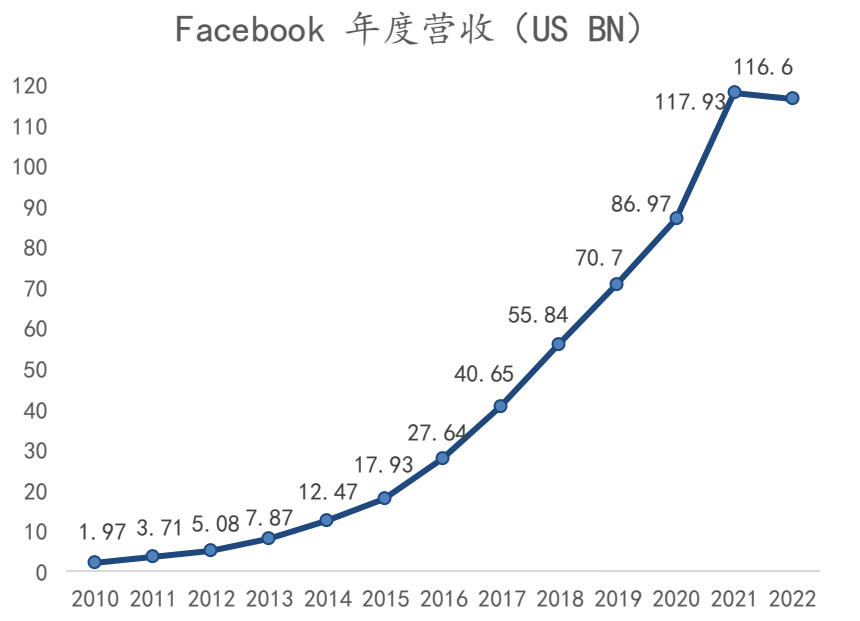
性加速，互联网软件如 Facebook 的收入增长，硬件如 Cisco 计算机基础设施；

图 43 达成全球 1 亿产品所需时间（年）



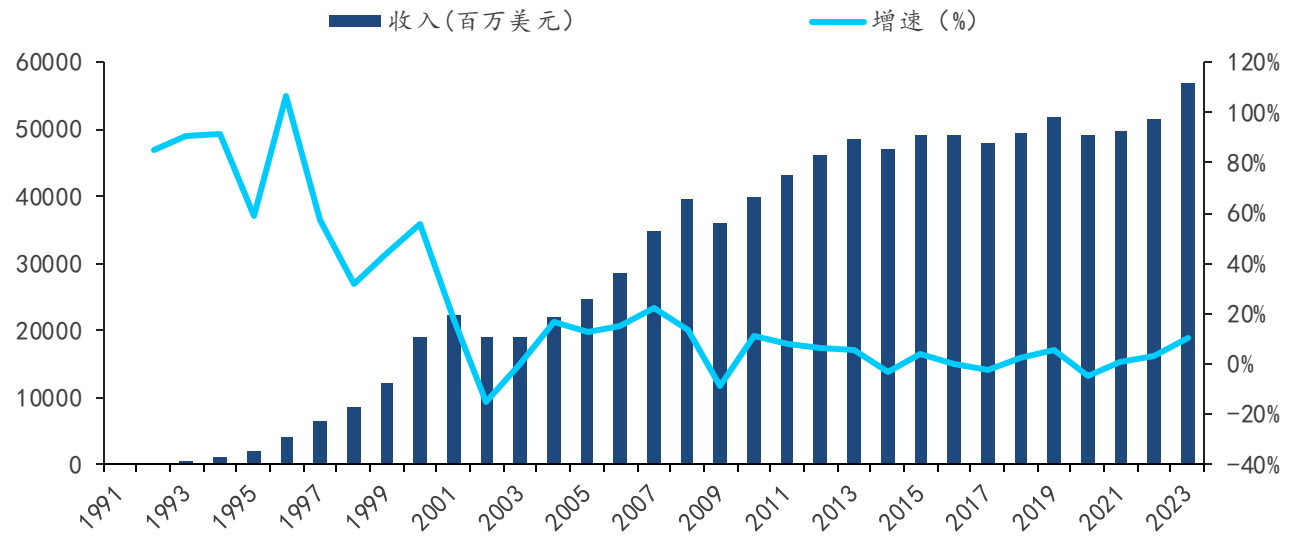
数据来源：格隆汇，国泰君安证券研究

图 44 Facebook 的收入呈现超线性增长



数据来源：Wind，国泰君安证券研究

图 45 Cisco 的收入呈现超线性增长



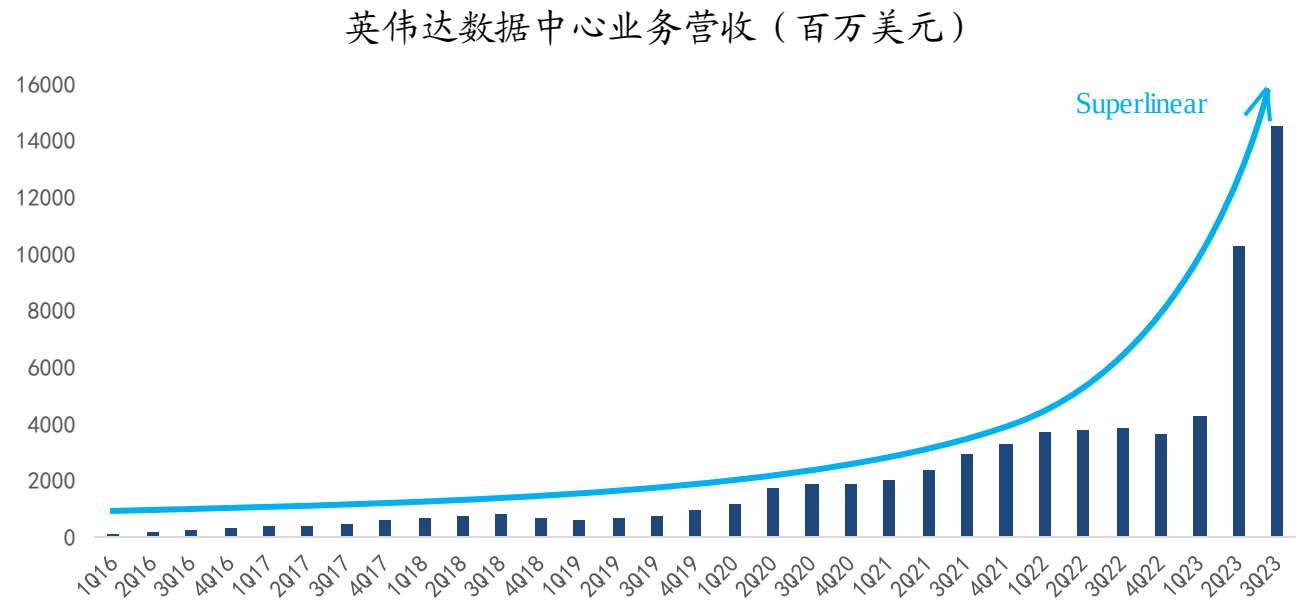
数据来源：Wind，国泰君安证券研究

超线性往往呈现出马太效应，在一个维度的探索达到一定阈值开启另外维度的增长，x^n， n＞1,本质上代表多维强势因子的叠加，如英伟达的 GPU 达到业界极限后，开启了 AI 超算，CPU,网络等产品超线性增长和发展：

超线性增长带来加速的增长曲线形态，在竞争中远远甩开竞争对手，竞争中胜出者往往受损较小，能够较大概率获得下一个周期较好成长：

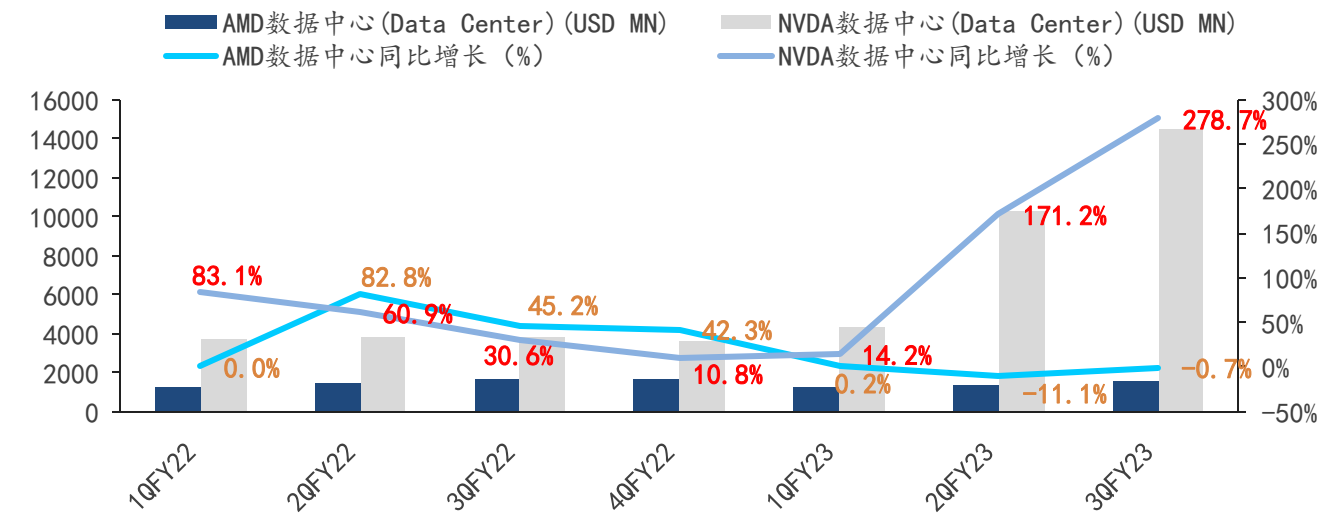
英伟达 24Q1 数据中心收入 42.8 亿美金，24Q2 为 103.2 亿美金，145.1 亿美金，我们惊人的发现 24 年 Q1+Q2 的收入约等于 Q3 的收入，呈现出明显的阶梯递增，阈值突破；英伟达从极致的 GPU 表现扩展到 AI 芯片，CPU，网络设备，以 278.7%的数据中心业务对比 AMD16 亿美金（ -0.7%）同比季度增速，在远期 2027 年很大可能实现 3000 亿美金的数据中心系统整体收入，英伟达以 2 年一代芯片架构的自我系统演进速度，正高速进入超线性增长旅程。

图 46 NVDA 的数据中心收入将呈现超线性增长



数据来源：NVDA，国泰君安证券研究

图 47 NVDA&AMD 的数据中心收入竞争对比



数据来源：NVDA，AMD，国泰君安证券研究

5. 风险提示

1. 大模型技术成熟速度及安全规范控制不及预期；
2. 地缘政治限制芯片自由贸易；
3. 技术行业可能会受到宏观经济周期的影响。在经济衰退期间，技术产品和服务的需求可能会下降；
4. 作为一家国际公司，英伟达可能会受到不同国家和地区法律法规变化的影响，尤其是与贸易政策和技术出口相关的法规；
5. 英伟达作为一家技术公司，特别是在显卡和 AI 领域，可能会受到快速技术变革和强烈竞争的影响等。

国泰君安海外科技团队介绍

深耕全球互联网，辐射海外大科技，全面覆盖社交、游戏、电商、互联网金融、互联网服务、AI 及硬科技、美股等领域，致力于结合产业视角与买方视角做差异化研究。

秦和平

执业证书编号：S0880123010042

海外科技领域负责人、首席分析师

梁昭晋

执业证书编号：S0880523010002

海外科技分析师

李奇

执业证书编号：S0880523060001

海外科技分析师

本公司具有中国证监会核准的证券投资咨询业务资格

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，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，本报告清晰准确地反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响，特此声明。

免责声明

本报告仅供国泰君安证券股份有限公司（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本公司的当然客户。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。

本报告的信息来源于已公开的资料，本公司对该等信息的准确性、完整性或可靠性不作任何保证。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌。过往表现不应作为日后的表现依据。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保持在最新状态。同时，本公司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。

本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。在任何情况下，本公司、本公司员工或者关联机构不承诺投资者一定获利，不与投资者分享投资收益，也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。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

本公司利用信息隔离墙控制内部一个或多个领域、部门或关联机构之间的信息流动。因此，投资者应注意，在法律许可的情况下，本公司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或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。在法律许可的情况下，本公司的员工可能担任本报告所提到的公司的董事。

市场有风险，投资需谨慎。投资者不应将本报告作为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向专业人士咨询并谨慎决策。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息或进而交易本报告中提及的证券。本报告不构成本公司向该机构之客户提供的投资建议，本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

评级说明

投资建议的比较标准
投资评级分为股票评级和行业评级。 以报告发布后的 12 个月内的市场表现为比较标准，报告发布日后的 12 个月内的公司股价（或行业指数）的涨跌幅相对同期的当地市场指数涨跌幅为基准。

	评级	说明
股票投资评级	增持	相对当地市场指数涨幅 15%以上
	谨慎增持	相对当地市场指数涨幅介于 5%～15%之间
	中性	相对当地市场指数涨幅介于-5%～5%
	减持	相对当地市场指数下跌 5%以上
行业投资评级	增持	明显强于当地市场指数
	中性	基本与当地市场指数持平
	减持	明显弱于当地市场指数

国泰君安证券研究所

	上海	深圳	北京
地址	上海市静安区新闸路 669 号博华广场 20 层	深圳市福田区益田路 6003 号荣超商务中心 B 栋 27 层	北京市西城区金融大街甲 9 号 金融街中心南楼 18 层
邮编	200041	518026	100032
电话	（021）38676666	（0755）23976888	（010）83939888

E-mail：gtjaresearch@gtjas.com

附：海外当地市场指数

亚洲指数名称	美洲指数名称	欧洲指数名称	澳洲指数名称
沪深 300	标普 500	希腊雅典 ASE	澳大利亚标普 200
恒生指数	加拿大 S&P/TSX	奥地利 ATX	新西兰 50
日经 225	墨西哥 BOLSA	冰岛 ICEX	
韩国 KOSPI	巴西 BOVESPA	挪威 OSEBX	
富时新加坡海峡时报		布拉格指数	
台湾加权		西班牙 IBEX35	
印度孟买 SENSEX		俄罗斯 RTS	
印尼雅加达综合		富时意大利 MIB	
越南胡志明		波兰 WIG	
富时马来西亚 KLCI		比利时 BFX	
泰国 SET		英国富时 100	
巴基斯坦卡拉奇		德国 DAX30	
斯里兰卡科伦坡		葡萄牙 PSI20	
		芬兰赫尔辛基	
		瑞士 SMI	
		法国 CAC40	
		英国富时 250	
		欧洲斯托克 50	
		OMX 哥本哈根 20	
		瑞典 OMXSPI	
		爱尔兰综合	
		荷兰 AEX	
		富时 AIM 全股	