全球科技动态追踪●计算机行业　　2024 年 3 月 28 日

GPT-5 有望今年夏季发布，多模态能力预期提升

计算机行业

推荐　　维持评级

摘要 | 03.15-03.22

- **股指动态：美股走强，中概股及港股小幅回落。**标普 500 指数+2.29%，纳斯达克综合指数+2.85%，费城半导体指数+3.16%；TAMAMA 科技指数+3.77%；纳斯达克中国金龙指数-2.78%；恒生科技指数-2.65%；计算机+3.51%。
- **个股表现：热门科技股大部分上涨。**据统计，相比 3 月 15 日收盘价，22 日盘后，苹果合计-0.20%，英伟达+7.35%，特斯拉+4.44%，谷歌+6.79%，亚马逊+2.55%，META+5.26%，微软+2.96%，ARM+5.65%，英特尔-0.16%，高通+1.73%，AMD-5.97%。
- **10 年期国债及汇率：**周内，中国 10 年期国债利率下降至 2.31%，累计下跌 1.49bps；美国 10 年期国债利率下降至 4.22%，累计下跌 9bps。3 月 22 日，美元兑人民币中间价报 7.10；较 3 月 15 日价累计调升 29 个基点。
- **核心观点**

近日，多家媒体公开消息称 GPT-5 预计将在今年夏季正式发布，目前仍处内测阶段。当前，最新版本 GPT-4 Turbo 已能支持最高 12.8 万 tokens 的输入，而谷歌近期发布的 Gemini 1.5 Pro 模型已经在输入长度方面实现了显著的突破，可支持 100 万 tokens 的输入，上下文输入长度方面大幅赶超。**鉴于此，我们预计，GPT-5 将在大模型的上下文输入长度实现重大突破，意味着 GPT-5 将有能力处理更长的文本，从而在理解和生成更复杂的语言结构上展现出更强大的能力。另外，新版本预计持续突破 GPT-4 的多模态能力，不仅限于处理文本、语音、图像等类型的信息，处理和理解多种类型的数据也将更加灵活，生成质量预计实现飞跃。迈向 AGI 的进程中，“超长文本”的处理能力和“超强模拟物理运动”的能力将被视为关键。GPT-5 的预期发布预计成为人工智能 AGI 发展重要催化剂，相关产业链投资机会依然凸显。**

- **风险提示**

技术迭代不及预期风险；科技巨头竞争加剧风险；法律监管风险；供应链风险；下游需求不及预期风险。

分析师

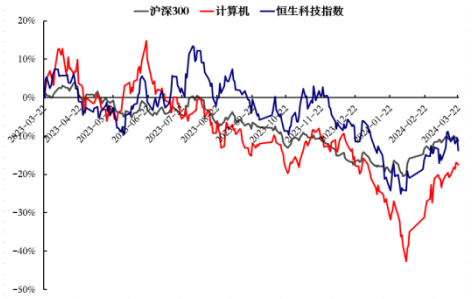
吴砚靖

☎：（8610）66568589

✉：wuyanjing@chinastock.com.cn

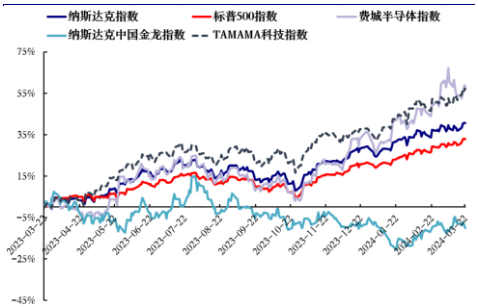
分析师登记编码：S0130519070001

国内表现　　2024-03-22



资料来源：Wind，中国银河证券研究院

全球行情　　2024-03-22



资料来源：Wind，中国银河证券研究院

相关研究

【银河计算机】全球科技动态追踪_Figure 联合 OpenAI 发布人形机器人，加速 AGI 时代到来

【银河计算机】全球科技动态追踪_Claude-3 AI 模型发布，博通 2024Q1 财报 AI 增长强劲

【银河计算机】全球科技动态追踪_超微电脑纳入标普 500 指数，新一轮 AI+硬件创新潮将至

目 录

一、全球市场表现

（一）股市动态

美股走强，中概股及港股小幅回落：标普 500 指数+2.29%，纳斯达克综合指数+2.85%，费城半导体指数+3.16%；TAMAMA 科技指数+3.77%；纳斯达克中国金龙指数-2.78%；恒生科技指数-2.65%；计算机+3.51%。

表 1：主要股指周变动

指数代码	指数简称	涨跌幅%					市盈率
		本周	上周	本月	本年度	2023	PE（TTM）
SPX.GI	标普500指数	2.29	-0.13	2.71	9.74	24.23	26.02
IXIC.GI	纳斯达克指数	2.85	-0.70	2.09	9.44	43.42	41.81
SOX.GI	费城半导体指数	3.16	-4.04	3.84	17.55	64.90	50.95
8884057.WI	TAMAMA科技指数	3.77	0.64	3.67	14.66	67.81	37.09
HXC.GI	纳斯达克中国金龙指数	-2.78	3.83	-1.78	-5.27	-3.39	19.59
HSTECH.HI	恒生科技指数	-2.65	4.85	0.72	-8.19	-8.83	21.08
CI005027.WI	计算机	3.51	2.07	8.61	-1.90	8.90	89.47

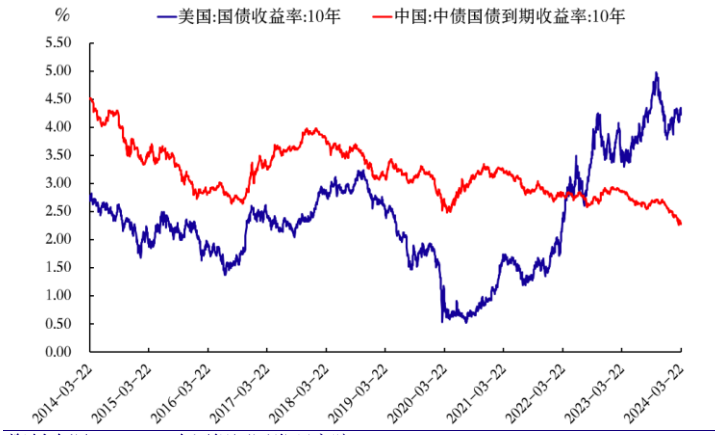
资料来源：Wind，中国银河证券研究院

（二）债市及汇率情况

债市：周内，中国 10 年期国债利率下降至 2.31%，累计下跌 1.49bps；美国 10 年期国债利率下降至 4.22%，累计下跌 9bps。

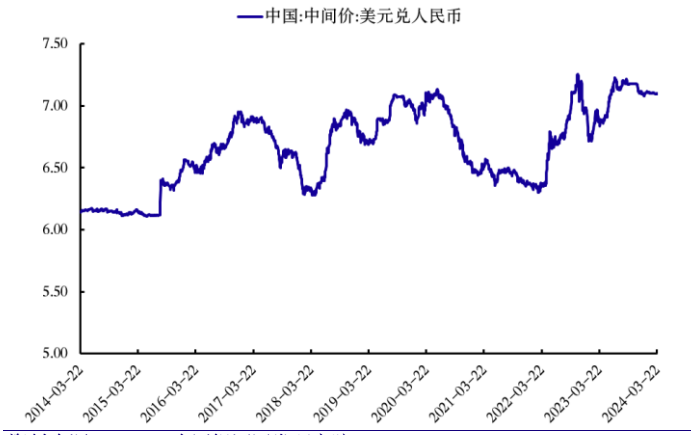
汇率：3 月 22 日，美元兑人民币中间价报 7.10；较 3 月 15 日价累计调升 29 个基点。

图1：国债收益率（10 年期）



资料来源：Wind，中国银河证券研究院

图2：美元兑人民币汇率(中间价)



资料来源：Wind，中国银河证券研究院

（三）重点公司表现

热门科技股大部分上涨：据统计，相比 3 月 15 日收盘价，22 日盘后，苹果合计-0.20%，英伟达+7.35%，特斯拉+4.44%，谷歌+6.79%，亚马逊+2.55%，META+5.26%，微软+2.96%，ARM+5.65%，英特尔-0.16%，高通+1.73%，AMD-5.97%。

表 2：重点公司周数据

所属板块	地区	证券代码	公司名称	股价（美元/港元/新台币）		区间涨跌幅（%）	总市值（亿美元/亿港元/亿新台币）[2024-03-22]	市销率 PS（TTM）	市盈率 PE（TTM）	PE(2022)	PE(2023)	PE(2024E)
				[2024-03-15]	[2024-03-22]							
算力及终端	美股	AAPL.O	苹果(Apple)	172.62	172.28	-0.20	26,603.27	6.90	26.36	20.71	30.87	25.82
		NVDA.O	英伟达(NVIDIA)	878.37	942.89	7.35	23,572.25	38.69	79.21	36.86	280.04	83.95
		TSLA.O	特斯拉(TESLA)	163.57	170.83	4.44	5,440.58	5.62	36.28	30.98	52.67	46.48
		HPQ.N	惠普(HP)	30.42	30.05	-1.22	294.03	0.55	8.61	8.24	9.14	0.00
		CSCO.O	思科(CISCO)	48.93	49.78	1.74	2,015.69	3.52	15.00	16.57	16.28	0.00
		ASML.O	阿斯麦	940.21	979.96	4.23	3,866.82	12.64	44.45	36.70	34.34	47.47
		AMD.O	超威半导体(AMD)	191.06	179.65	-5.97	2,902.76	12.80	339.90	79.12	278.85	68.43
		INTC.O	英特尔(INTEL)	42.64	42.57	-0.16	1,799.86	3.32	106.56	13.61	125.43	42.48
		QCOM.O	高通(QUALCOMM)	167.20	170.10	1.73	1,898.32	5.23	24.45	9.51	22.36	19.41
		ARM.O	ARM	126.97	134.15	5.65	1,379.16	46.94	1,622.54	0.00	147.03	0.00
		ON.O	安森美半导体(ON SEMICONDUCTOR)	74.87	74.68	-0.25	319.13	3.87	14.61	14.18	16.47	16.86
	港股	0909.HK	明源云	2.32	2.52	8.62	48.95	2.68	-7.57	-10.65	-8.66	-15.87
		9698.HK	万国数据-SW	7.51	8.01	6.66	122.11	1.14	-8.25	-21.99	0.00	-7.24
		1686.HK	新意网集团	2.68	2.55	-4.85	59.65	2.35	6.57	11.66	7.98	0.00
	台股	2330.TW	台积电	753.00	785.00	4.67	228,988.26	9.42	24.28	11.71	18.34	20.02
		2454.TW	联发科	1,135.00	1,125.00	-0.88	17,995.76	4.15	23.38	8.46	21.09	0.00
互联网	美股	GOOGL.O	谷歌(ALPHABET)-C	141.18	150.77	6.79	18,803.45	6.10	25.40	19.04	23.69	22.06
		AMZN.O	亚马逊(AMAZON)	174.42	178.87	2.55	18,579.91	3.23	61.07	-314.82	51.61	43.96
		META.O	脸书(META PLATFORMS)	484.10	509.58	5.26	12,991.26	9.63	33.23	13.75	23.27	26.52
		NFLX.O	奈飞(NETFLIX)	605.88	628.01	3.65	2,717.77	8.06	50.25	29.21	39.40	37.80
		PDD.O	拼多多	123.74	122.99	-0.61	1,634.03	4.67	19.28	22.77	22.94	13.29
		NTES.O	网易	106.93	105.52	-1.32	687.55	4.66	16.38	16.34	14.46	15.75
		BIDU.O	百度	103.86	102.18	-1.62	352.24	1.89	12.49	36.42	14.52	11.75
		TCOM.O	携程网	43.42	45.02	3.68	310.10	4.90	21.98	116.72	17.58	24.11
		BABA.N	阿里巴巴	73.42	72.13	-1.76	1,849.71	1.40	13.01	23.78	18.64	12.05
	港股	9988.HK	阿里巴巴-SW	71.90	71.00	-1.25	14,465.71	1.41	13.11	23.80	18.53	12.12
		0700.HK	腾讯控股	283.80	288.80	1.76	27,322.94	4.06	21.49	15.17	21.90	15.71
		80700.HK	腾讯控股-R	261.40	268.60	2.75	24,807.05	4.17	22.06	0.00	21.86	16.05
		9999.HK	网易-S	170.10	166.80	-1.94	5,377.00	4.71	16.56	16.52	13.96	15.90
		9888.HK	百度集团-SW	101.30	98.20	-3.06	2,754.73	1.85	12.29	36.49	14.48	11.45
		89888.HK	百度集团-SWR	93.40	91.25	-2.30	2,501.07	1.90	12.60	0.00	14.50	11.72
		1024.HK	快手-W	50.65	49.00	-3.26	2,131.82	1.69	30.21	-19.96	32.60	15.97
		81024.HK	快手-WR	46.70	45.55	-2.46	1,935.52	1.73	30.98	0.00	32.58	16.22
		9626.HK	哔哩哔哩-W	92.90	86.75	-6.62	365.44	1.47	-6.87	-8.90	-7.41	-17.60
		2518.HK	汽车之家-S	52.00	50.85	-2.21	259.02	3.27	12.49	14.79	12.86	11.79
		9898.HK	微博-SW	76.00	70.15	-7.70	165.04	1.20	6.16	52.13	7.52	6.02
软件及应用	美股	MSFT.O	微软(MICROSOFT)	416.42	428.74	2.96	31,857.25	14.00	38.60	24.58	38.62	35.75
		SNOW.N	SNOWFLAKE	156.97	159.03	1.31	523.69	18.66	-62.63	-67.89	-82.25	0.00
		ORCL.N	甲骨文(ORACLE)	125.54	127.79	1.79	3,512.33	6.69	33.00	32.81	34.08	26.18
		CRM.N	赛富时(SALESFORCE)	294.33	307.77	4.57	2,985.37	8.56	72.18	91.82	1224.61	0.00
		ADBE.O	奥多比(ADOBE)	492.46	499.52	1.43	2,260.56	11.34	47.09	32.69	50.01	40.05
		INTU.O	财捷(INTUIT)	625.52	643.74	2.91	1,802.34	11.94	65.07	52.92	73.39	64.42
		SNPS.O	新思科技(SYNOPSYS)	550.03	594.20	8.03	906.41	14.79	64.40	49.43	63.63	0.00
		CDNS.O	铿腾电子(CADENCE)	298.44	322.74	8.14	879.79	21.51	84.50	51.91	71.17	0.00
		ADSK.O	欧特克(AUTODESK)	254.24	262.86	3.39	562.30	10.23	62.06	81.13	63.29	0.00
		U.N	Unity	26.09	26.99	3.45	104.17	4.76	-12.61	-9.35	-18.80	0.00
	港股	0020.HK	商汤-W	0.84	0.80	-4.76	267.75	6.45	-4.11	-10.98	0.00	-8.95
		80020.HK	商汤-WR	0.78	0.75	-3.85	243.10	6.56	-4.18	0.00	0.00	-9.23
		3888.HK	金山软件	22.15	25.00	12.87	334.79	3.41	62.76	-5.25	61.69	27.73
		0268.HK	金蝶国际	9.15	9.40	2.73	337.79	5.29	-145.84	-133.49	-178.41	-912.73
		9878.HK	汇通达网络	29.80	30.40	2.01	171.02	0.19	37.41	70.68	0.00	27.70
		3650.HK	KEEP	4.01	3.99	-0.50	20.97	0.88	2.33	0.00	0.00	-8.17
		0354.HK	中国软件国际	5.28	4.96	-6.06	144.19	0.71	24.67	24.18	0.00	11.23
		1357.HK	美图公司	3.15	3.50	11.11	156.72	5.14	37.54	57.75	38.62	26.60
		3896.HK	金山云	1.67	1.83	9.58	69.64	0.90	-2.90	-2.69	-3.18	4.50
		2013.HK	微盟集团	2.12	2.05	-3.30	57.29	2.46	-3.16	-8.29	0.00	-44.40
		1675.HK	亚信科技	6.88	7.50	9.01	70.15	0.80	11.93	13.73	13.69	6.67
		2121.HK	创新奇智	6.59	6.73	2.12	38.03	1.91	-10.78	-31.65	0.00	0.00
		2400.HK	心动公司	15.08	16.82	11.54	80.81	2.07	-96.46	-16.75	0.00	24.18
		0777.HK	网龙	11.94	12.36	3.52	65.66	0.82	7.87	9.71	0.00	5.86

资料来源：Wind，中国银河证券研究院

二、行业要闻

（一）算力及终端

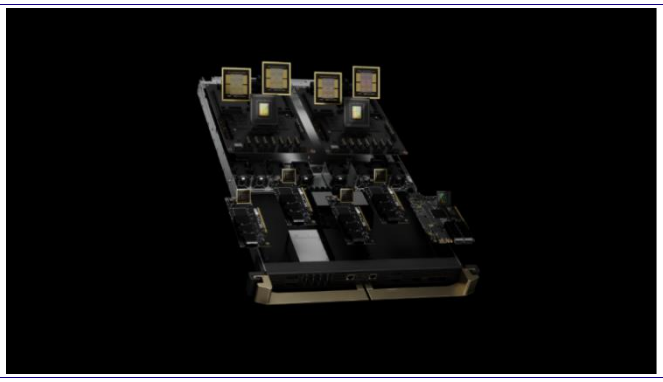
【NVIDIA Blackwell 平台发布，赋能计算新时代】

2024 年 3 月 18 日英伟达宣布推出 NVIDIA Blackwell 架构。全新 GB200 GPU 有 2080 亿个晶体管，采用台积电 4NP 节点工艺，提供 20 petaflops FP4 的算力，与 H100 相比，B200 的晶体管数量是其 2 倍多，而单个 H100 最多提供 4 petaflops 算力，B200 比其提升了 5 倍的性能。

GB200 Grace Blackwell 超级芯片是通过 NVLink-Chip-to-Chip（C2C）以 900GB/s 的超低功耗将两个 Blackwell NVIDIA B200 Tensor Core GPU 与一个 Grace CPU 相连而成。

GB200 包含两个 Grace CPU 和四个 Blackwell GPU。GB200 装有液冷连接口，支持高速网络的 PCIe gen 6，以及具有用于 NVLink 电缆盒的 NVLink 连接器。GB200 可以提供 80petaflops 的 AI 性能和 1.7 TB 的快速内存。

图3：英伟达 GB200 GPU



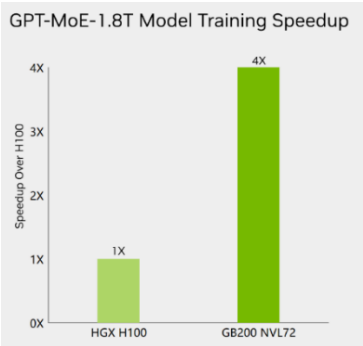
资料来源：Nvidia 官网，中国银河证券研究院

为扩大 Blackwell 的规模，英伟达构建了一款名为 NVLink Switch 的新芯片。每个芯片可以以每秒 1.8TB 的速度连接四个 NVLink，并通过减少网络内流量来消除流量拥塞。

英伟达另外推出 GB200 NVL72，一套多节点液冷机架扩展系统（liquid-cooled, rack-scale architecture），内置 NVIDIA BlueField-3 数据处理器，可在超大规模 AI 云中实现云网络加速、组合式存储。由 8 个 NVL72 机架系统组合在一起就是 DGX GB800，有 288 个 Grace CPU、576 个 BlackwellGPU，240TB 的内存和 11.5 exaflop FP4 计算。该平台可作为一个单 GPU，具有 1.4 exaflops 的 AI 性能和 30TB 的快速内存。通过 NVIDIA Quantum-X800 InfiniBand 网络和 NVIDIA Spectrum™-X800 网络，这一系统可以扩展到数万个 GB200 芯片，组成最新一代 DGX SuperPOD 的基础模块。GB200 NVL72 应用了第五代 NVLink，可在单个 NVLink 域中连接多达 576 个 GPU，总带宽超过 1PB/s 和 240TB 的快速内存。每个 NVLink 交换机托盘提供 144 个 100GB 的 NVLink 端口。

大规模训练上，GB200 NVL72 内含的第二代 Transformer 引擎，具有 FP8 精度，**可将大型语言模型的训练速度提高 4 倍。**

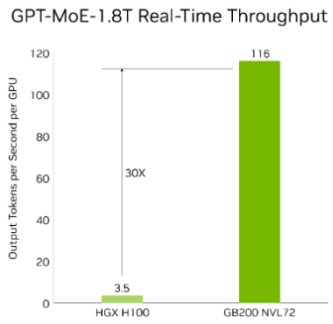
图4：大型语言模型的训练速度提高 4 倍



资料来源：Nvidia 官网，中国银河证券研究院

AI 推理方面，GB200 NVL72 具有第二代 Transformer 引擎，可加速 LLM 推理工作负载。**与上一代 H100 相比，其为 1.8T 参数 GPT-MoE 等资源密集型应用加速了 30 倍。**

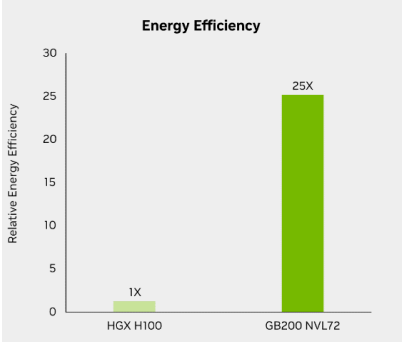
图5：与 H100 相比，GB200 的实时吞吐量提高了 30 倍



资料来源：Nvidia 官网，中国银河证券研究院

节能方面，液冷 GB200 NVL72 机架系统可减少数据中心的碳足迹和能耗。液体冷却可提高计算密度，减少使用的占用面积，并促进与大型 NVLink 域架构的高带宽、低延迟 GPU 通信。**与 NVIDIA H100 风冷基础设施相比，GB200 在相同功耗下提供 25 倍的性能，同时减少用水量。**

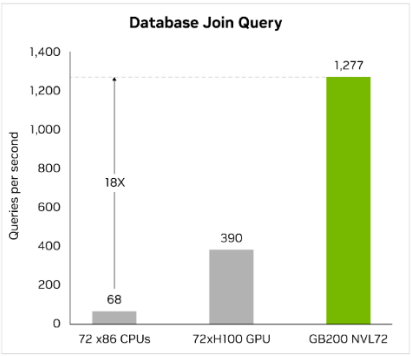
图6：GB200 在与 H100 相同功耗下提供 25 倍的性能



资料来源：Nvidia 官网，中国银河证券研究院

数据处理方面，为了在 GPU 上高效处理数据集，Blackwell 架构引入了硬件解压缩引擎，该引擎可以大规模本地解压压缩数据，并进行端到端加速。该解压引擎支持 LZ4、Deflate 和 Snappy 压缩格式的数据。解压缩引擎加快了内存绑定的操作，提供高达 800GB/s 的速度，并使 Grace Blackwell 的执行速度比 CPU（Sapphire Rapids）快 18 倍，比 NVIDIA H100 Tensor Core GPU 快 6 倍。凭借 8TB/s 的高内存带宽和 Grace CPU 高速 NVlink-Chip-to-Chip（C2C），该引擎加快了数据库查询的整个过程。**使得用户能够快速获得数据，同时降低成本。**

图7：GB200NVL72、72x H100 和 72x86CPU 的数据库联接查询吞吐量

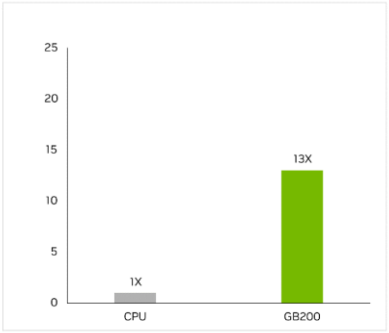


资料来源：Nvidia 官网，中国银河证券研究院

基于物理模拟方面，基于物理的仿真模拟仍然是产品设计和开发的支柱。从飞机和火车到桥梁、硅芯片，甚至药品，通过仿真测试和改进产品可节省数十亿美元。专用集成电路几乎完全在 CPU 上设计，工作流程复杂，包括使用模拟分析来识别电压和

电流。以 Cadence SpectreX 仿真器为具体例子，SpectreX 在 GB200 上的运行速度比在 x86 CPU 上快 13 倍。

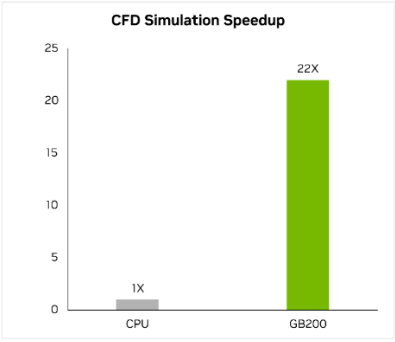
图8：Cadence SpectreX 模拟器在 GB200 上的运行速度比在 x86 CPU 上快 13 倍



资料来源：Nvidia 官网，中国银河证券研究院

业界普遍使用 GPU 用于计算流体动力学（CFD）。工程师和设备设计师使用它来研究和预测他们的设计。以 Cadence Fidelity 大型涡流模拟器（LES）为例，在 GB200 上运行模拟的速度比 x86 CPU 快 22 倍。

图9：Cadence Fidelity 在 GB200 上运行仿真的速度比 x86CPU 快 22 倍



资料来源：Nvidia 官网，中国银河证券研究院

【英伟达推出 Blackwell 架构 DGX SuperPOD】

英伟达于 2024 年 3 月 18 日发布新一代 AI 超级计算机—搭载 NVIDIA GB200 Grace Blackwell 超级芯片的 NVIDIA DGX SuperPOD，可以用于处理万亿参数模型，能够保证超大规模生成式 AI 训练和推理工作负载的持续运行。

DGX SuperPOD 采用新型高效液冷机架级扩展架构（liquid-cooled rack-scale architecture），基于 NVIDIA DGX GB200 系统搭建，在 FP4 精度下可提供 11.5 exaflops 的 AI 超级计算性能和 240TB 的快速显存，且可通过增加机架来扩展性能。每个 DGX GB200 系统搭载 36 个 GB200 超级芯片，共包含 36 个 Grace CPU 和 72 个 Blackwell GPU，通过第五代 NVLink 连接成一台超级计算机。这些芯片通过 NVIDIA Quantum InfiniBand 网络连接，可扩展到数万个 GB200 超级芯片，且每块 GPU 可以达到每秒 1800GB 的带宽。

DGX SuperPOD 每台超级计算机都在出厂前完成了搭建、布线和测试，从而加快了在用户数据中心的部署速度。

DGX SuperPOD 具有智能预测管理功能，能够持续监控软硬件中的数千个数据点，通过预测并拦截可能导致停机和低效的因素来节省时间、能耗和计算成本。即使没有系统管理员在场，该软件也能识别需要重点关注的领域并制定维护计划，灵活调整计算资源，通过自动保存和恢复作业来防止停机。如果软件检测到需要更换组件，该软

件将激活备用容量以确保工作能够及时完成。为硬件更换做好安排，以免出现计划之外的停机。

英伟达发布了一款统一用于 AI 模型训练、微调和推理的通用 AI 超级计算平台 NVIDIA DGX B200 系统。DGX B200 是 DGX 系列的第六代产品，包含 8 个 NVIDIA B200 Tensor Core GPU 和 2 个第五代英特尔 Xeon 处理器。

DGX SuperPOD 也可以使用 DGX B200 系统构建。DGX B200 系统凭借 Blackwell 架构中的 FP4 精度，可提供 144 petaflops 的 AI 性能、1.4TB GPU 显存和 64TB/s 的显存带宽，从而使得该系统的万亿参数模型实时推理速度比上一代产品提升了 15 倍。DGX B200 系统包含带有 8 个 NVIDIA ConnectX-7 网卡和 2 个 BlueField-3 DPU，每个连接的带宽达到 400Gb/s，可通过 NVIDIA Quantum-2 InfiniBand 和 NVIDIA Spectrum-X 以太网网络平台支持更高的 AI 性能。

软件方面，所有 NVIDIA DGX 平台均包含用于企业级开发和部署的 NVIDIA AIEnterprise 软件。DGX 用户可以通过使用该软件平台中的预训练的 NVIDIA 基础模型、框架、工具套件和 NVIDIA NIM 微服务来加速工作。

NVIDIA 将在今年晚些时候提供基于 DGX GB200 和 DGX B200 系统构建而成的 NVIDIA DGX SuperPOD。

图10：NVIDIA DGX SuperPOD

资料来源：Nvidia 官网，中国银河证券研究院

【英伟达发布 GR00T 人形机器人基础模型和 Isaac 机器人平台更新】

英伟达于 2024 年 3 月 18 日发布人形机器人通用基础模型 Project GR00T。除此之外，还发布了一款基于 NVIDIA Thor 系统级芯片（SoC）的新型人形机器人计算机 Jetson Thor，并对 NVIDIA Isaa 机器人平台进行了重大升级，包括生成式 AI 基础模型和仿真工具，以及 AI 工作流基础设施。

GR00T 驱动的机器人将能够理解自然语言，并通过观察人类行为来模仿动作、快速学习协调、灵活性和其它技能，以适应现实世界并与之互动。

Jetson Thor 是一个全新的计算平台，能够执行复杂的任务，并安全、自然地与人和机器交互。该 SoC 包括一个带有 Transformer Engine 的 GPU，采用了 Blackwell 架构，可提供每秒 800 万亿次 8 位浮点运算 AI 性能，以运行 GR00T 等多模态生成式 AI 模型。凭借集成的功能的安全处理器、高性能 CPU 集群和 100GB 带宽，简化了设计和集成工作。

Isaac 平台也有重大更新，GR00T 使用的 Isaac 工具能够为在任何环境中的任何机器人创建新的基础模型。这些工具包括用于强化学习的 Isaac Lab 和 OSMO。新的 Isaac

Lab 是一个 GPU 加速、性能优化的轻量级应用，基于 Isaac Sim 而构建，专门用于运行数千个用于机器人学习的并行仿真。

英伟达还发布了 Isaac Manipulator 和 Isaac Perceptor 等一系列机器人预训练模型、库和参考硬件。Isaac Manipulator 为机械臂提供了灵活性和模块化 AI 功能，并提供了基础模型和 GPU 加速库。它提供了 80 倍的路径规划加速，零样本感知提高了效率和吞吐量，使开发者能够实现更多新的机器人任务的自动化。Isaac Perceptor 提供了多摄像头和 3D 环绕视觉功能，这些功能正在被制造业和物流业中的自主移动机器人所采用，以提高效率和更好地保护工人，同时降低错误率和成本。

新的 Isaac 平台功能将在下个季度推出。

（二）大模型及云应用

【多家媒体消息称 Open AI 预计将于今年夏季发布 GPT-5】

OpenAI 预计在未来几个月内发布其 ChatGPT 模型的下一个版本。据知情人士透露，可能在夏季发布 GPT-5。另一位则表示，一些企业客户最近已经看到了最新模型及其对 ChatGPT 工具的相关增强演示。

OpenAI 目前还没有为新模型设置具体的发布日期，这意味着目前内部预期可能改变。据知情人士透露，OpenAI 仍在训练 GPT-5。训练完成后，它将在内部进行安全测试，并进一步进行“红队测试”，红队测试通常由外部专家或组织内部专门的红队人员执行，该测试的目的是在发布前找到问题。红队安全测试需要完成的具体时间框架尚不明确，因此这一过程可能会推迟任何发布日期。

当前，最新版本 GPT-4 Turbo 已能支持最高 12.8 万 tokens 的输入，而谷歌近期发布的 Gemini 1.5 Pro 模型已经在输入长度方面实现了显著的突破，可支持 100 万 tokens 的输入，上下文输入长度方面大幅赶超。**鉴于此，我们预计，GPT-5 将在大模型的上下文输入长度实现重大突破，意味着 GPT-5 将有能力处理更长的文本，从而在理解和生成更复杂的语言结构上展现出更强大的能力。另外，新版本预计持续突破 GPT-4 的多模态能力，不仅限于处理文本、语音、图像等类型的信息，处理和理解多种类型的数据也将更加灵活，生成质量预计实现飞跃。迈向 AGI 的进程中，“超长文本”的处理能力和“超强模拟物理运动”的能力将被视为关键。GPT-5 的预期发布预计成为人工智能 AGI 发展重要催化剂，相关产业链投资机会依然凸显。**

【甲骨文和英伟达将在全球范围内提供主权人工智能】

甲骨文和英伟达宣布扩大合作，为全球客户提供主权人工智能解决方案。英伟达称，主权人工智能指一个国家利用自己的基础设施、数据、人力资源和商业网络来生产人工智能的能力。甲骨文的分布式云、AI 基础设施和生成式 AI 服务与英伟达的加速计算和生成式 AI 软件相结合，支持政府和企业部署 AI 工厂。这些 AI 工厂可以在本地，也可以在一个国家或机构的安全场所内运行云服务，并具有一系列操作控制功能，从而支持实现多样化和促进经济增长的目标。

英伟达 A 的全栈 AI 平台与甲骨文的 Enterprise AI 相结合，为客户提供 AI 解决方案，可以更好地控制运营、位置和提高安全性，以帮助满足数字主权要求。**在 26 个国家的 66 个云区域，客户可以跨基础设施和应用程序访问 100 多个云和 AI 服务，以支持 IT 迁移、现代化和创新。**

两家公司的组合产品可以通过公共云部署，也可以部署在特定位置的客户数据中心，并具有灵活的操作控制。OCI 服务和定价在各种部署中是一致的，以简化规划、移植和管理。

甲骨文的云服务利用了一系列英伟达堆栈，包括英伟达加速计算基础设施和 NVIDIA AI Enterprise 软件平台。

为了帮助客户满足日益增长的 AI 模型需求，甲骨文计划在 OCI Supercluster 和 OCI Compute 上利用最新 NVIDIA Grace Blackwell 计算平台。通过使用新的 OCI Compute 裸机实例、低延迟 RDMA 网络和高性能存储，OCI Supercluster 将大幅提速。OCI Compute 将采用 NVIDIA GB200 Grace Blackwell 超级芯片和 NVIDIA Blackwell B200 Tensor Core GPU。

【英伟达推出 CUDA-X 微服务，助力企业迈向生成式 AI】

NVIDIA AI Enterprise 5.0 于 2024 年 3 月 18 日发布，其中包括英伟达 CUDA-X 微服务、用于部署生成式 AI 应用和加速计算的可下载软件容器。

NVIDIA AI Enterprise 5.0 包括一系列微服务，其中包括用于部署 AI 模型的 NVIDIA NIM 和包括 NVIDIA cuOpt 在内的 NVIDIA CUDA-X 微服务集合。

NIM 微服务可为来自英伟达及其合作伙伴生态系统的数十种热门 AI 模型优化推理。NIM 由英伟达推理软件（包括 Triton 推理服务器、TensorRT 和 TensorRT-LLM）提供支持，可将部署时间由数周缩短至数分钟。

NVIDIA cuOpt 是一款 GPU 加速的 AI 微服务，能够支持动态决策，从而帮助企业降低成本、缩短时间并减少碳足迹。例如 NVIDIA RAG LLM operator 将把 Co-Pilot 和其他使用检索增强生成的生成式 AI 应用从试验阶段推向生产阶段，而无需重写任何代码。包括 CrowdStrike、SAP 和 ServiceNow 在内的多家领先应用和网络安全平台提供商都正在采用英伟达微服务。

NVIDIA AI Enterprise 5.0 还集成了一个开发者工具包，叫做 NVIDIA AI Workbench，用于快速下载、自定义和运行生成式 AI 项目。该软件目前已全面上市，并已获得 NVIDIA AI Enterprise 许可证支持。此外，5.0 版本现在还支持红帽 OpenStack 平台，大多数财富 500 强公司都使用该平台创建私有云和公有云服务。由红帽负责维护，为开发者构建虚拟计算环境提供熟悉的平台.

【英伟达发布 Omniverse Cloud API，为工业数字孪生软件工具提供助力】

英伟达于 2024 年 3 月 18 日宣布将以 API 形式提供 Omniverse Cloud，将工业数字孪生应用和工作流创建平台的覆盖范围扩展至整个软件制造商生态系统。借助五个 Omniverse Cloud 应用编程接口（API），开发者能够将 Omniverse 的核心技术直接集成到现有的数字孪生设计与自动化软件应用中，或是集成到用于测试和验证机器人或自动驾驶汽车等自主机器的仿真工作流中。

五个全新 Omniverse Cloud API 既可单独使用，也可组合使用。分别是：1）USD Render：生成 OpenUSD 数据的全光线追踪 NVIDIA RTX 渲染；2）USD Write：用户能够修改 OpenUSD 数据并与之交互；3）USD Query：支持场景查询和交互式场景；4）USD Notify：追踪 USD 变化并提供更新信息；5）Omniverse Channel：连接用户、工具和世界，实现跨场景协作。

（三）计算机设备

【英伟达发布全新交换机，全面优化万亿参数级 GPU 计算和 AI 基础设施】

英伟达于 2024 年 3 月 18 日发布专为大规模 AI 订制的网络交换机-X800 系列。NVIDIA Quantum-X800 InfiniBand 网络和 NVIDIA Spectrum-X800 网络是全球首批达到 800Gb/s 端到端吞吐量的网络平台。与其配套软件配合可进一步加速数据中心中的 AI、云、数据处理和高性能计算(HPC)应用

Quantum-X800 平台包含了 NVIDIA Quantum Q3400 交换机和 NVIDIA ConnectX-8 SuperNIC，二者互连可以达到端到端 800Gb/s 吞吐量，交换带宽容量较上一代产品提高了 5 倍，网络计算能力凭借英伟达的 SHARP 技术(SHARPv4)提高了 9 倍，达到了 14.4Tflops。

Spectrum-X800 平台可以优化 AI 云和企业级基础设施的网络性能。借助 800Gb/s 的 Spectrum SN5600 交换机和 NVIDIA BlueField-3 SuperNIC，Spectrum-X800 平台可为多租户生成式 AI 云和大型企业级用户提供各种至关重要的功能。Spectrum-X800 通过优化网络性能，加快 AI 工作负载的处理、分析和执行速度，进而缩短 AI 解决方案的开发、部署和上市时间。Spectrum-X800 专为多租户环境打造，实现了每个租户的 AI 工作负载的性能隔离，使业务性能能够持续保持在最佳状态，提升客户满意度和服务质量。

全球多家头部基础设施供应商和系统厂商将在明年开始提供基于 Quantum-X800 和 Spectrum-X800 的网络平台，包括 Aivres、DDN、戴尔科技、Eviden、Hitachi Vantara、慧与、联想、超微和 VAST Data 等。

三、风险提示

技术迭代不及预期风险；科技巨头竞争加剧风险；法律监管风险；供应链风险；下游需求不及预期风险。

图表目录

表格目录

分析师承诺及简介

本人承诺以勤勉的执业态度，独立、客观地出具本报告，本报告清晰准确地反映本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与、未来也将不会与本报告的具体推荐或观点直接或间接相关。

吴砚靖，TMT/科创板研究负责人。北京大学软件项目管理硕士，10 年证券分析从业经验，历任中银国际证券首席分析师，国内大型知名 PE 机构研究部执行总经理。具备一二级市场经验，长期专注科技公司研究。

免责声明

本报告由中国银河证券股份有限公司（以下简称银河证券）向其客户提供。银河证券无需因接收人收到本报告而视其为客户。若您并非银河证券客户中的专业投资者，为保证服务质量、控制投资风险、应首先联系银河证券机构销售部门或客户经理，完成投资者适当性匹配，并充分了解该项服务的性质、特点、使用的注意事项以及若不当使用可能带来的风险或损失。

本报告所载的全部内容只提供给客户做参考之用，并不构成对客户的投资咨询建议，并非作为买卖、认购证券或其它金融工具的邀请或保证。客户不应单纯依靠本报告而取代自我独立判断。银河证券认为本报告资料来源是可靠的，所载内容及观点客观公正，但不担保其准确性或完整性。本报告所载内容反映的是银河证券在最初发表本报告日期当日的判断，银河证券可发出其它与本报告所载内容不一致或有不同结论的报告，但银河证券没有义务和责任去及时更新本报告涉及的内容并通知客户。银河证券不对因客户使用本报告而导致的损失负任何责任。

本报告可能附带其它网站的地址或超级链接，对于可能涉及的银河证券网站以外的地址或超级链接，银河证券不对其内容负责。链接网站的内容不构成本报告的任何部分，客户需自行承担浏览这些网站的费用或风险。

银河证券在法律允许的情况下可参与、投资或持有本报告涉及的证券或进行证券交易，或向本报告涉及的公司提供或争取提供包括投资银行业务在内的服务或业务支持。银河证券可能与本报告涉及的公司之间存在业务关系，并无需事先或在获得业务关系后通知客户。

评级标准

评级标准		评级	说明
评级标准为报告发布日后的 6 到 12 个月行业指数（或公司股价）相对市场表现，其中：A 股市场以沪深 300 指数为基准，新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准，北交所市场以北证 50 指数为基准，香港市场以摩根士丹利中国指数为基准。	行业评级	推荐：	相对基准指数涨幅 10%以上
		中性：	相对基准指数涨幅在-5%～10%之间
		回避：	相对基准指数跌幅 5%以上
	公司评级	推荐：	相对基准指数涨幅 20%以上
		谨慎推荐：	相对基准指数涨幅在 5%～20%之间
		中性：	相对基准指数涨幅在-5%～5%之间
		回避：	相对基准指数跌幅 5%以上

联系

中国银河证券股份有限公司研究院

深圳市福田区金田路 3088 号中洲大厦 20 层

上海浦东新区富城路 99 号震旦大厦 31 层

北京市丰台区西营街 8 号院 1 号楼青海金融大厦

公司网址：www.chinastock.com.cn

机构请致电：

深广地区：程曦 0755-83471683chengxi_yj@chinastock.com.cn
苏一耘 0755-83479312suyiyun_yj@chinastock.com.cn

上海地区：陆韵如 021-60387901luyunru_yj@chinastock.com.cn
李洋洋 021-20252671liyangyang_yj@chinastock.com.cn

北京地区：田薇 010-80927721tianwei@chinastock.com.cn
唐嫚羚 010-80927722tangmanling_bj@chinastock.com.cn