行业深度报告● 电子行业 2023 年 03 月 27 日

AI 商业模式逐步落地，算力产业链迎接星辰大海

核心观点：

- **AI 带动万亿蓝海市场，“模型+数据+算力”为产业核心壁垒。**AI 历史发展余 70 年，当前正处于新一轮产业变革制高点。从规模上看，全球 AI 产业规模预计 2030 年将达到 1500 亿，未来 8 年复合增速约 40%。从市场来看，美国领先，中国和欧盟并驾齐驱，三地企业合计份额为 70.01%。2022 年中国人工智能产业规模达 1958 亿元，年增长率 7.8%，整体稳健增长。模型、数据和算力为人工智能发展三驾马车，Transformer 模型的引入标志着自然语言处理模型能够大规模地生成类似人类的语言，并且进入可大规模、可复制的大工业落地阶段。算法模型发展的同时，对于数据规模和质量的要求也在不断提高。其中 ChatGPT 参数量达到 1750 亿次，数据规模达到 45TB，从 1956-2020 年，计算机处理能力的 FLOPS 增加了一万亿倍。同时海外和国内互联网行业巨头积极布局 AI，竞赛压力逐步提升。
- **AI 产业迎来“iPhone”时刻，英伟达召开 GTC2023，关注高性能计算相关领域壁垒。**英伟达在 **GTC2023** 推出 AI Foundations 云服务，从 NEMO、PICASSO、BIONEMO 三方面，赋能不同 AI 场景。同时发布 H100 NVL 服务器，相比 A100 DGX 提供 10 倍的计算速度。在大算力背景下，存算性能呈现剪刀差，存储器件性能远弱于算力性能提升，AI 训练未来的瓶颈不是算力，而是 GPU 的“内存墙”。因此，未来存算一体化趋势确定，HBM 与 Chiplet 有望实现降本增效，全球半导体厂商已提出多种解决方案，存内计算电路可基于 SRAM 和 NOR Flash 实现。HBM 的高带宽技术，基于 TSV 和芯片堆叠技术的堆叠可实现高于 256GBp/s 带宽远超过 DDR4 和 GDDR6。Chiplet 技术无需中介层、芯片直接通过 TSV 直接进行高密度互连,性能可以得到很大的提升，算力水平也会提高。
- **AI 商业落地曙光出现，ChatGPT 引爆大算力需求。**ChatGPT 是使用海量语料库进行训练的语言生成器,在 2022 年 11 月 ChatGPT 推出后，迅速引爆市场，2 个月内月活跃用户数便达一亿，成为了历史上用户增长最快的消费应用。ChatGPT 参数量 2018 年 OpenAI 发布的 ChatGPT 1.0 的模型参数为 1.17 亿，2019 年的第二代模型参数为 15 亿，ChatGPT 3.0 的参数相比于 ChatGPT2.0 增长了近百倍，达到了 1750 亿。强大的算力水平是 AI 大模型必备的技术支撑, ChatGPT 3.0 模型需要使用 1024 颗英伟达 A100 芯片训练长达一个月的时间,AIGC 商业落地蓄势待发，未来对算力的需求更将超乎想象。

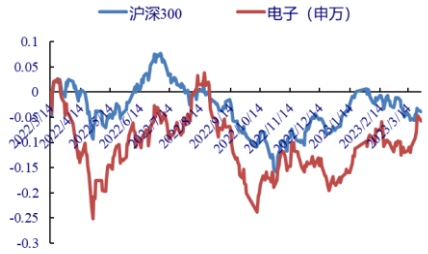
电子行业

推荐(维持)

分析师
高峰
☎：010-80927671
✉：gaofeng_yj@chinastock.com.cn
分析师登记编码：S0130522040001

王子路
☎：010-80927632
✉：wangzilu_yj@chinastock.com.cn
分析师登记编码：S0130522050001

行业相对沪深 300 表现图



资料来源：Wind，中国银河证券研究院

相关研究

【银河电子】行业月报_功率半导体高景气有望延续，集成电路静待周期回暖
【银河电子】电子行业年度策略报告_以自主可控为基，以创新成就未来 221213
【银河电子】行业深度报告_电子行业三季报总结_三季度业绩承压，行业底部信号初现

- **我们认为，在 ChatGPT 等应用商业化出现落地方式，AIGC 创作内容不断增长的条件下，芯片作为 AI 行业的基础设计，为 AI 训练和数据计算提供支持，未来 AI 应用落地层面对庞大算力的需求更为重要，因此，相关算力产业链未来发展值得期待。**我们看好国内相关算力产业链公司的未来发展，建议关注：GPU、加速卡、AI 芯片：寒武纪（688256.SH）、景嘉微（300474.SZ）、海光信息（688041.SH ）;先进封装：通富微电（002156.SZ ）、长电科技（600584.SH）、深科技（000021.SZ）;服务器及加速卡 PCB：沪电股份（002463.SZ ）、胜宏科技（300476.SZ）;AIoT 产业链：瑞芯微（603893.SH ）、全志科技（300458.SZ ）、晶晨股份（688099.SH）、富瀚微（300613.SZ）;芯片 IP：芯原股份（688521.SH ）、华大九天（301269.SZ）;存储芯片/模组/PCIe：兆易创新（603986.SH ）、江波龙（301308.SZ ）、北京君正（300223.SZ）、聚辰股份（688123.SH ）、澜起科技（688008.SH）;散热材料：中石科技（300684.SZ ）、飞荣达（300602.SZ）。

建议关注

股票名称	股票代码	当前价格	EPS(元)			PE（X）		
			2022E	2023E	2024E	2022E	2023E	2024E
寒武纪-U	688256.SH	183.00	-1.89	-1.09	-	-	-	-
景嘉微	300474.SZ	113.05	0.63	0.93	1.27	178.12	121.55	89.10
海光信息	688041.SH	66.25	0.61	0.91	-	107.88	72.49	-
通富微电	002156.SZ	24.79	0.39	0.74	1.06	63.08	33.40	23.45
长电科技	600584.SH	33.40	1.85	2.04	2.37	18.10	16.40	14.07
深科技	000021.SZ	15.34	0.53	0.66	0.74	28.81	23.17	20.81
沪电股份	002463.SZ	21.60	0.88	1.12	1.42	24.50	19.34	15.21
胜宏科技	300476.SZ	19.50	0.99	1.22	1.54	19.76	15.94	12.62
瑞芯微	603893.SH	95.85	0.93	1.46	2.07	103.48	65.67	46.41
全志科技	300458.SZ	28.90	0.54	0.69	-	53.47	42.10	-
晶晨股份	688099.SH	88.38	2.78	3.81	-	31.77	23.20	-
富瀚微	300613.SZ	74.31	1.85	2.40	3.12	40.24	30.90	23.81
华大九天	301269.SZ	122.34	0.34	0.47	0.64	355.33	259.25	190.92
兆易创新	603986.SH	117.90	3.72	4.03	4.98	31.68	29.22	23.67
江波龙	301308.SZ	83.77	0.96	1.38	1.89	87.31	60.70	44.44
北京君正	300223.SZ	85.43	1.94	2.44	3.16	44.15	35.05	27.07
聚辰股份	688123.SH	103.93	4.70	6.17	-	22.12	16.84	-
澜起科技	688008.SH	70.17	1.64	2.37	-	42.75	29.65	-
中石科技	300684.SZ	18.97	0.69	0.97	1.35	27.52	19.56	14.07
飞荣达	300602.SZ	17.35	0.07	0.40	0.80	250.00	43.09	21.75

资料来源：Wind，中国银河证券研究院

目　　录

一、 AI 带动万亿蓝海市场，“模型+数据+算力”为产业核心壁垒

（一）AI 产业历经 70 年发展，终将迎来第四次工业革命

AI 历史发展余 70 年，经历多发展阶段，最早可追溯至上世纪初期。AI 目前已渗透至日常生活方方面面，在医疗保健、汽车、金融、游戏、环境监测、农业、体育、能源管理、安全等各个领域的大量应用正在改变人类的生活、工作和娱乐方式。这些技术的进一步发展将迎来第四次工业革命。,造成这一现象的原因包括计算机技术的进步(高性能计算、网格和云计算)、代码共享度提高（GitHub、GitLab、BitBucket 等服务）以及大量开源软件。AI 将为企业和国家经济系统提供革命改变，商业领域，人工智能带来的优势包括：快速揭示大数据中的模式、快速进行可视化和分析、改进产品设计等等，并进一步有望提升服务水平、增加利润、扩大业务、提高效率和成本结构。

表 1： 人工智能发展历程

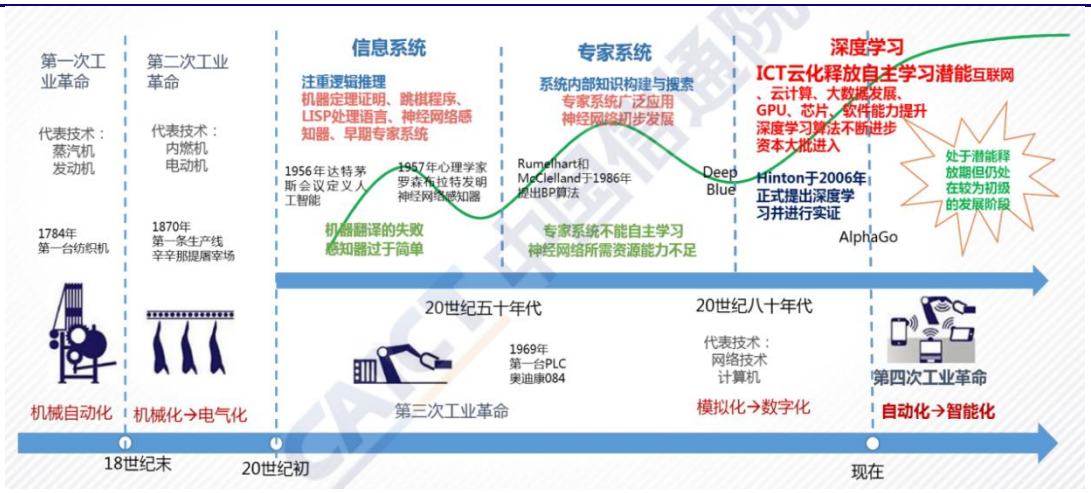
发展阶段	年份	发展历程
萌芽	1900-1956	1900 年，希尔伯特在数学家大会上宣布了 23 个未解决的问题，其中第二和第十个问题与人工智能密切相关，最终促进了计算机的发明。1954 年，冯-诺依曼完成了早期计算机 EDVAC 的设计，并提出了 "冯-诺依曼架构"。图灵、哥德尔、冯-诺依曼、维纳、克劳德-香农和其他的先驱者奠定了人工智能和计算机技术的基础。
黄金时代	1956-1974	1965 年，麦卡锡、明斯基等科学家召开 "达特茅斯会议"，首次提出 "人工智能（AI）"的概念，标志着人工智能学科的诞生。随后，人工智能研究进入了 20 年的黄金时代，取得了一批令人瞩目的研究成果，如机器定理证明和跳棋程序，掀起了人工智能发展的第一个高潮。在这个黄金时代，约翰-麦卡锡开发了 LISP 语音，成为此后几十年人工智能领域最主要的编程语言；马文-明斯基对神经网络有了更深入的研究，也发现了简单神经网络的缺点；接着开始出现多层神经网络和反向传播算法。
第一次寒冬	1974-1980	人工智能发展的最初突破极大地提高了人们的期望，使人们高估了科技发展的速度。然而，连续的失败和预期目标的落空使人工智能的发展进入低谷。1973 年，赖特-希尔关于人工智能的报告，拉开了人工智能冬天的序幕。此后，科学界对人工智能进行了一轮深入的拷问，使人工智能受到了严厉的批评和对其实用价值的质疑。随后，政府和机构也停止或减少了资助，人工智能在 20 世纪 70 年代陷入了它的第一个冬天。有限的计算能力和大量常识性数据的缺乏使发展陷入瓶颈，尤其是过度依赖计算能力和经验数据量的神经网络技术，在很长一段时间内没有取得实质性的进展。
应用发展	1980-1987	专家系统模拟人类专家的知识和经验来解决特定领域的问题，实现了人工智能从理论研究到实际应用的重大突破。专家系统在医学、化学、地质学等领域的成功，将人工智能推向了应用发展的新高潮，1980 年 XCON 在卡内基梅隆大学（CMU）正式启动，成为专家系统开始在特定领域发挥作用的里程碑，推动了整个人工智能技术进入繁荣阶段。经过十年的沉寂，神经网络有了新的研究进展，并发现了具有学习能力的神经网络算法，这使得神经网络的发展在 20 世纪 90 年代后期一路走向商业化，被应用于文字图像识别和语音识别。

第二次寒冬	1987-1993	随着人工智能应用规模的不断扩大，应用领域狭窄、缺乏常识性知识、知识获取困难、推理方法单一、缺乏分布式功能、与现有专家系统数据库难以兼容等问题逐渐暴露出来。当时的人工智能领域主要使用约翰-麦卡锡的 LISP 编程语言。LISP 机的逐步发展被蓬勃发展的个人电脑打败了，专用 LISP 机的硬件销售市场严重崩溃，人工智能领域再次进入寒冬。硬件市场的崩溃和理论研究的混乱，再加上政府和机构纷纷停止对人工智能研究领域的资金投入，导致人工智能领域几年来一直处于低迷状态。但另一方面在理论方法的研究上也取得了一些成果。1988 年，美国科学家朱迪亚-皮尔将概率统计方法引入人工智能的推理过程；IBM 的沃森研究中心将概率统计方法引入到人工智能的语言处理中；1992 年，李开复利用统计方法设计开发了世界上第一个独立于扬声器的连续语音识别程序；1989 年，AT&T 贝尔实验室的亚恩-莱坤和团队将卷积神经网络技术应用在了人工智能的手写数字图像识别中。
稳步发展	1993-2011	人工智能的创新研究因网络技术的发展而加速，尤其是互联网的发展，使人工智能技术进一步实用化。1995 年，理查德-华莱士开发了新的聊天机器人程序 Alice，它能够利用互联网不断增加自己的数据集并优化内容。1997 年，IMB 的计算机 Deep blue 深蓝击败了世界象棋冠军卡斯帕罗夫。德国科学家霍克赖特和施米德赫伯提出了 LSTM 递归神经网络，至今仍被用于手写识别和语音识别，对后来的人工智能研究产生了深远影响。2004 年，美国神经科学家杰夫·霍金斯出版了《人工智能的未来》，2006 年，杰弗里辛顿出版了《学习多层表征》，为神经网络奠定了一个新的架构，对未来人工智能中的深度学习的研究产生了深刻影响。
深化阶段	2012-至今	随着移动互联网技术和云计算技术的爆发，积累了难以想象的数据量，为人工智能的后续发展提供了足够的素材和动力，以深度神经网络为代表的人工智能技术的快速发展，大大跨越了科学与应用之间的 "技术鸿沟"，迎来了爆发式增长。2012 年，多伦多大学在 ImageNet 视觉识别挑战赛上设计的深度卷积神经网络算法，被认为是深度学习革命的开始。2014 年，Ian Goodfellow 提出了 GANs 生成式对抗网络算法，这是一种用于无监督学习的人工网络。这是一种用于无监督学习的人工智能算法，由生成网络和评估网络组成，这种方法很快被人工智能的许多技术领域所采用。2016 年和 2017 年，谷歌推出的人工智能程序 AlphaGo 连续击败了前围棋世界冠军韩国的李世石，以及现任围棋世界冠军中国的柯洁，引起了巨大轰动。同时语音识别、图像识别、无人驾驶等技术不断进步。2022 年 11 月，OpenAI 推出其开发的一个人工智慧聊天机器人程序 ChatGPT。该程序使用基于 GPT-3.5 架构的大型语言模型并通过强化学习进行训练，成为 AIGC 现象级应用。

资料来源：《人工智能导论——人工智能的发展历史、现状及发展趋势》，银河证券研究院

当前正处于第四次工业革命的风口浪尖，正处于新一轮产业变革制高点。当下全球正在发生的第四次工业革命是人工智能、智慧网联时代，以超大数据、超强算力、超强算法的人工智能为核心技术，以智能家居、智能音箱、智慧城市、智能汽车和手机为数据入口的智能终端产品正加速 AI 时代的进化。

图 1：全球人工智能产业浪潮



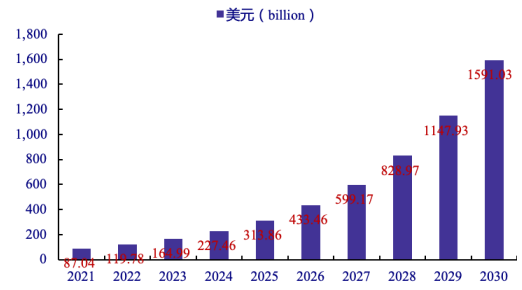
资料来源：中国信通院，中国银河证券研究院

（二）产业规模扩容厂商竞入新蓝海，国家政策 AI 发展

全球 AI 产业规模预计 2030 年将达到 1500 亿，未来 8 年复合增速约 40%。目前全球人工智能企业的数量迅速增长，2022 年，全球人工智能（AI）市场规模估计为 197.8 亿美元，预计到 2030 年将达到 1591.03 亿美元，从 2022 年到 2030 年，复合年增长率为 38.1%。

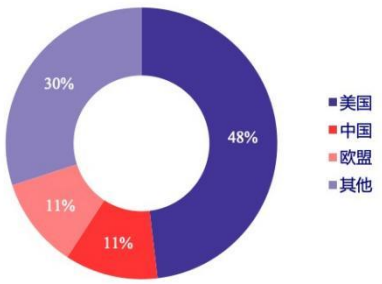
从地区上来看，美中欧暂时领先，格局仍未确定。其中 AI 市场美国领先，中国和欧盟并驾齐驱。截至 2017 年，全球人工智能企业主要集中在美国（2905 家）、中国（670 家）和欧盟（657 家），如图所示，合计份额为 70.01%。目前，美国仍是人工智能的核心发源地之一，其他国家也在迅速跟进人工智能的研发。国内北京人工智能发展领跑全国，上海、广东、江苏、浙江等地发展逐渐加快。

图 2： 人工智能全球市场规模预测



资料来源：Precedence Research，中国银河证券研究院

图 3： 2017 年人工智能全球产业格局



资料来源：《人工智能产业化的历史、现状与发展趋势》，中国银河证券研究院

图 4：人工智能全球产业分布

资料来源：中国信通院，中国银河证券研究院

2022 年中国人工智能产业规模达 1958 亿元，年增长率 7.8%，整体稳健增长。而从应用格局来看，机器视觉、智能语音和自然语言处理是中国人工智能市场规模最大的三个应用方向。根据清华大学数据显示，三者占比分别为 34.9%、24.8%和 21%。一方面，政策推动下国内应用场景不断开放，各行业积累的大量数据为技术落地和优化提供了基础条件。另一方面，以百度、阿里、腾讯和华为为代表的头部互联网和科技企业加快在三大核心技术领域布局，同时一系列创新型独角兽企业在垂直领域快速发展，庞大的商业化潜力推动核心技术创新。

图 5：中国人工智能产业规模

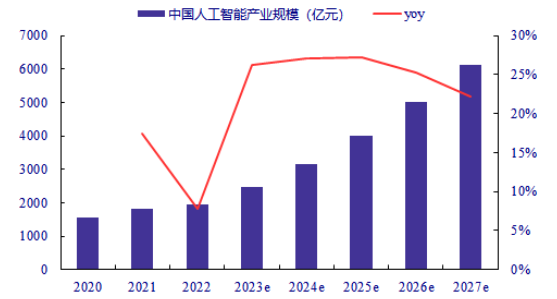
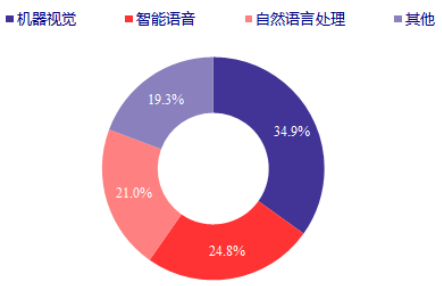


图 6：人工智能应用方向



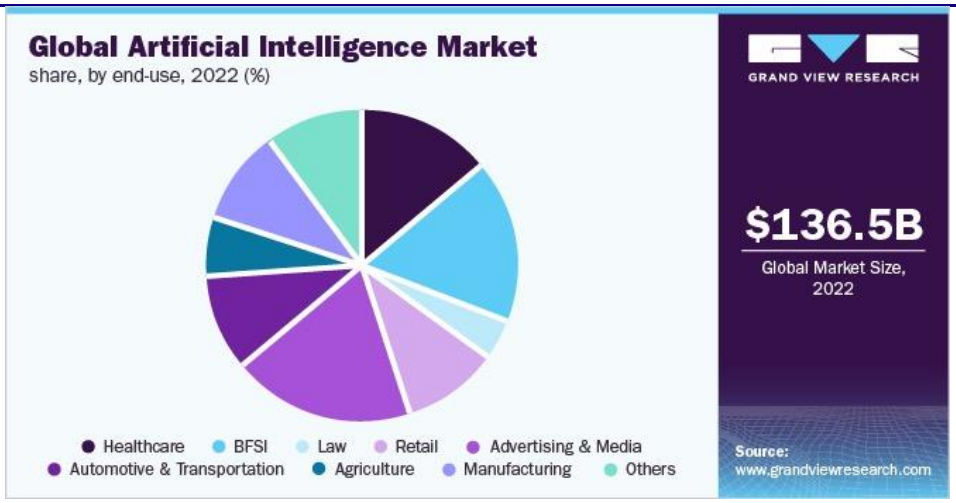
资料来源：艾瑞咨询，中国银河证券研究院

AI 场景丰富，多垂直细分领域均有应用。

科技的持续研究和创新正在推动人工智能技术在行业垂直领域的应用，如汽车、医疗、零售、金融和制造业。例如，2020 年 11 月，英特尔公司收购了 Cnvrg.io，为数据科学家建立和运行机器学习模型开发和运营平台，以促进其人工智能业务。深度学习和 ANN（人工神经网络）的进步也推动了人工智能在航空航天、医疗保健、制造和汽车等多个行业的应用，Google 一直在采用 ANN 来改善路线，并处理使用 ANN 收到的反馈。计算机视觉技术的最新进步，如 GAN（Generative Adversarial Networks）和 SSD（Single Shot MultiBox Detector），已经促成了数字图像处理技术的诞生，这些技术可以使在低光或低分辨率下拍摄的图像和视频转换为高清质量，计算机视觉的持续研究为安全与监控、医疗保健和运输等部门的数字图像处理奠定了基础。

在人工智能不同的垂直应用领域中，广告和媒体部门引领市场在 2022 年占全球收入份额的 19.5%以上，这一高份额归因于人工智能营销应用程序不断地增长。预计到 2030 年，医疗保健部门将获得最大份额。基于机器人辅助手术、减少剂量错误、虚拟护理助理、临床试验参与者标识符、医院工作流程管理、初步诊断和自动图像诊断等用例，医疗保健部门已独树一帜。

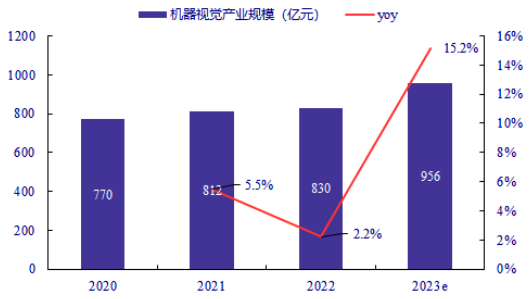
图 7： 人工智能应用市场细分



资料来源：Grand View Research，中国银河证券研究院

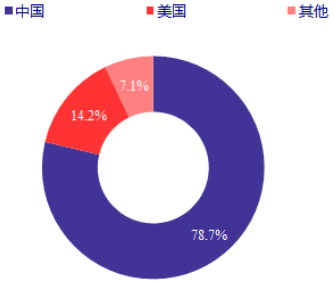
在国内，机器视觉领域是人工智能应用最多最广的板块。2022 年机器视觉相关投融资浪潮高企，工业、泛安防、能源赛道热度高涨，持续受到资本青睐。近两年来是 AI 产业上市最火爆的细分赛道，涌现了商汤科技、格灵深瞳、云从科技、奥比中光等 IPO 企业， 2022 年我国机器视觉产品的市场规模达到 830 亿元。同时，庞大的市场牵引科技研发，我国在全球机器视觉技术创新上已位居世界前列。截至 2021 年 8 月，中国机器视觉专利申请量占全球机器视觉专利总申请量的 78.7%；其次是美国，占比为 14.2%。

图 8：中国机器视觉产业规模



资料来源：艾瑞咨询，中国银河证券研究院

图 9：截至 2021 年 8 月机器视觉专利申请数



资料来源：艾瑞咨询，中国银河证券研究院

国家政策不断发力，助力中国 AI 成长。近年来，人工智能产业发展受到国家层面的重视，相关政策频出。2017 年，国务院出台《新一代人工智能发展规划》，成为中国人工智能发展的指导性文件；国家发改委、中央网信办、工信部等部门陆续发布人工智能相关细则，部署人工

智能发展计划。近五年来，中国政府凭借在人工智能产业发展中强有力的领导地位，发挥资源聚集的制度优势。国家坚持“市场导向”，秉持开源开放原则，在推动产学研用多主体共享成果的同时还加强军民深度融合，实现创新资源共享和科技成果双向转化，不断通过政策更好地引导人工智能产业全方位快速发展。

表 2：近五年中国人工智能政策

政策	时间	部门	政策内容
《2018 年政府工作报告》	2018.03	国务院	“人工智能”再次被写入政府工作报告，强调要加强新一代人工智能研发应用，在医疗、养老、教育、文化、体育等多领域推进 “互联网+”
《高等学校人工智能创新行动计划》	2018.04	教育部	聚焦并加强新一代人工智能基础理论和核心关键技术研究，加快建设人工智能科技创新基地，加快建设一流人才队伍和高水平创新团队
《机器人产业发展规划（2016—2020 年）》	2018.04	工信部、发改委、财政部	开展人工智能、机器人深度学习等基础前沿技术研究，围绕人工智能、感知与识别、机构与驱动，控制与交互等方面开展基础和共性关键技术研究
《关于发展数字经济稳定并扩大就业的指导意见》	2018.09	发改委	加快形成适应数字经济发展的就业政策体系，大力提升数字化、网络化、智能化就业创业服务能力，大力培育互联网，物联网、大数据、云计算、人工智能等领域的就业机会
《新一代人工智能产业创新点任务揭榜工作方案》	2018.11	国务院	征集并遴选一批掌握人工智能核心关键技术，创新能力强、发展潜力大的企业、科研机构等，调动产学研用各方积极性
《2019 年政府工作报告》	2019.03	国务院	促进新兴产业加快发展。深化大数据，人工智能等研发应用，培育新一代信息技术，高端装备、生物医药，新能源汽车、新材料等新兴产业集群，壮大数字经济。加快在各行业各领域推进“互联网+”
《关于促进人工智能和实体经济深度融合的指导意见》	2019.03	中央深改委	提出促进人工智能和实体经济深度融合，坚持以市场需求为导向，以产业应用为目标，深化改革创新，优化制度环境，激发企业创新活力和内生动力，结合不同行业、不同区域特点，探索创新成果应用转化的路径和方法，构建数据驱动，人机协同、跨界融合，共创分享的智能经济形态
《新一代人工智能治理原则》	2019.06	科技部	突出了发展负责任的人工智能这一主题，强调了和谐友好、公平公正、包容共享、尊重隐私。安全可控、共担责任、开放协作、敏捷治理等八条原则
《国家新一代人工智能创新发展试验区建设工作指引》	2019.08	科技部	提出开展人工智能技术应用示范、人工智能政策试验、人工智能社会实验，积极推进人工智能基础设施建设，到 2023 年，布局建设 20 个左右试验区
《中共中央、国务院关于构建更加完善的要素市场化配路体制机制的意见》	2020.04	中共中央、国务院	培育数字经济新产业、新业态和新模式，支持构建农业、工业、交通、教育、安防、城市管理、公共资源交易等领域规范化数据开发利用的场景。发挥行业协会商会作用，推动人工智能、可穿戴设备、车联网、物联网等领域数据采集标准化。
《2020 政府工作报告》	2022.05	国务院	推动制造业升级和新兴产业发展。大幅增加制造业中长期贷款。发展工业互联网，推进智能制造。电商网购、在线服务等新业态在抗疫中发挥

			了重要作用，要继续出台支持政策，全面推进“互联网+”打造数字经济新优势。
《国家新一代人工智能标准体系建设指南》	2022.08	工信部等五部委	为加强人工智能领域标准化顶层设计，推动人工智能产业技术研发和标准制定，促进产业健康可持续发展，国家标准化管理委员会、中央网信办、国家发展改革委、科技部及工业和信息化部印发《国家新一代人工智能标准体系建设指南》，AI 产业迎来顶层设计。
《国家新一代人工智能创新发展试验区建设工作指引》	2022.09	客人基本	鼓励直辖市、副省级城市、地级市等地方申请建设国家新一代人工智能创新发展试验区，科技部将从政策、资源等方面对试验区建设给予支持。
《十四五规划和 2035 年远景目标》	2021.03	中共中央	围绕总体目标，规划纲要在三个方面布局人工智能发展。突破核心技术，打造数字经济新优势，营造良好数字环境。
《新一代人工智能伦理规范》	2021.09	科技部	提出了增进人类福祉、促进公平公正、保护隐私安全、确保可控可信、强化责任担当、提升伦理素养等 6 项基本伦理要求。同时，提出人工智能管理、研发、供应、使用等特定活动的 18 项具体伦理要求。
《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》	2022.07	科技部等六部门	以促进人工智能与实体经济深度融合为主线，以推动场景资源开放、提升场景创新能力为方向，强化主体培育、加大应用示范、创新体制机制、完善场景生态，加速人工智能技术攻关、产品开发和产业培育，探索人工智能发展新模式新路径，以人工智能高水平应用促进经济高质量发展。
《关于支持建设新一代人工智能示范应用场景的通知》	2022.08	科技部	坚持面向世界科技前沿、 面向经济主战场、面向国家重大需求、面向人民生命健康，充分发挥人工智能赋能经济社会发展的作用，围绕构建全链条、过程的人工智能行业应用生态，支持一批基础较好的人工智能应用场景，强研发上下游配合与新技术集成，打造形成一批可复制、可推广的标杆型示范应用场景。首批支持建设十个示范应用场景。
《质量强国建设纲要》	2023.02	中共中央、国务院	加快大数据、网络、人工智能等新技术的深度应用，促进现代服务业与先进制造业、现代农业融合发展。

资料来源：政府信息网，中国银河证券研究院

（三）人工智能发展三驾马车——模型、数据和算力

人工智能产业链按照上下游可以分为人工智能基础层、人工智能技术层、人工智能应用层。其中，上游人工智能基础层将 AI 分为模型、算力和数据三大要素。AI 模型生产工具包括 AI 算法框架、AI 开放平台、AI 开发平台和预训练模型；AI 算力基础领域包括 AI 芯片、智能服务器和云服务；AI 数据资源包括 AI 基础数据服务和数据治理。人工智能技术层包括计算机视觉、智能语音、自然语言处理、知识图谱、机器学习。人工智能应用层则很广泛，涵盖“AI+泛安防”、“AI+泛互联网”、人机交互、自主无人系统、“AI+媒体”、“AI+金融”、“AI+医疗”、“AI+工业”、“AI+零售”、“AI+政务”等应用，涉及经济社会运行的方方面面。

图 10：中国人工智能产业图谱

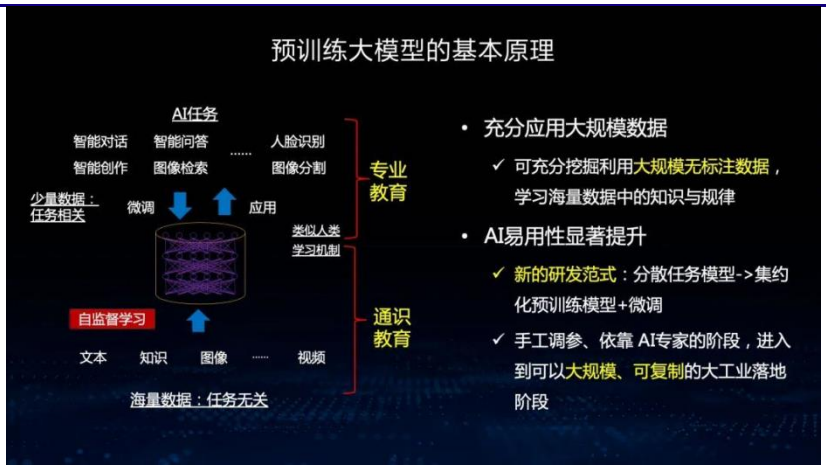
资料来源：艾瑞咨询，中国银河证券研究院

1、模型

人工智能框架一直在蓬勃发展，各种框架在开发者的不断开发和自然选择的基础上不断迭代。经过激烈的竞争，最终出现了双雄并立的 TensorFlow 和 PyTorch 的两大阵营。随后，迁移学习（Transfer learning）成为开发大规模人工智能模型的流行技术，使研究人员能够利用预先训练的模型来提高新任务的性能。在此期间，注意力机制（Attention mechanisms）也出现了，允许模型有选择地关注输入数据的某些部分。

2017 年，Transformer 模型的引入标志着自然语言处理的重大突破，使模型能够大规模地生成类似人类的语言。预训练大模型的基本原理是充分利用大规模的数据，以挖掘数据中的知识和规律，类似接受人类的通识教育。再针对特定的任务，进行参数微调，可以达到智能对话、智能问答、智能创作、人脸识别等功能，并且进入可大规模、可复制的大工业落地阶段。在算法模型层面，超大规模模型成为近几年来最热门的发展之一。

图 11： 预训练大模型基本原理



资料来源：AI 大模型公众号，中国银河证券研究院

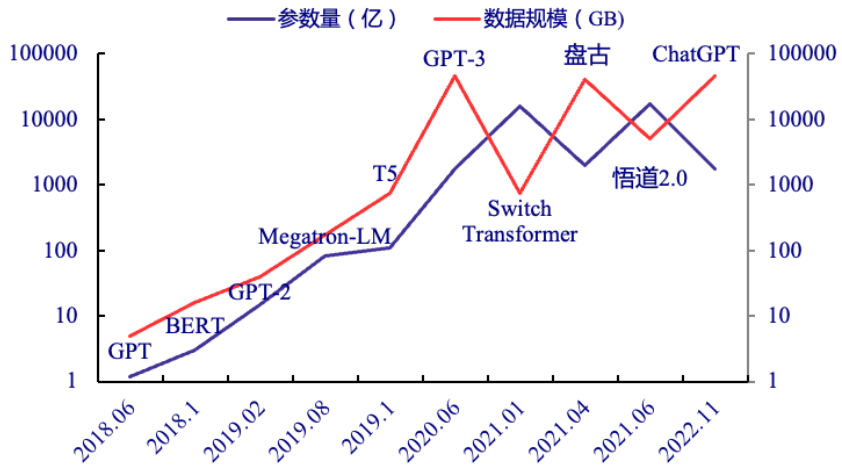
2018 年 OpenAI 推出了非常强大的预训练语言模型 Generative Pre-trained Transformer（GPT），结果表明这一模型可以在非常复杂的 NLP 任务中取得非常惊艳的效果，而且并不需要有监督学习进行模型微调。同年，谷歌的雅各布·德夫林和同事创建并发布了 BERT（Bidirectional Encoder Representations from Transformers）。BERT 是一个双向 transformer 模型，用于对大量未标记的文本数据进行预训练，以学习一种语言表示形式，这种语言表示形式可用于对特定机器学习任务进行微调。虽然 BERT 在几项任务中的表现都优于 NLP 领域沿用过的最先进的技术，但其性能的提高还是归功于双向 transformer、掩蔽语言模型对任务的训练以及结构预测功能，还包括大量的数据和谷歌的计算能力。此后，基于 BERT 的改进模型包括 DistillBERT、XLNet、RoBERTa、T5 等大量新式预训练语言模型不断涌现。

2019 年，OpenAI 继续推出了带有 15 亿参数的 GPT-2，这一改进后的模型能够生成连贯的文本段落，做到初步的阅读理解、机器翻译等。接着，英伟达推出了具有 83 亿参数的 Megatron-LM，谷歌继续推出了具有 110 亿参数的 T5 模型，微软推出了 170 亿参数的图灵 Turing-NLG。

2020 年，OpenAI 又推出了超大规模的语言训练模型 GPT-3，参数量达到了 1750 亿之高，实现了模型参数从亿级到上千亿的跨越。此后，谷歌、华为、阿里巴巴和北京智源等企业和研究机构纷纷推出超大规模的预训练模型，包括 MT-NLG、Switch Transformer、盘古和悟道 2.0 等。预训练的模型参数数量和训练数据量正以每年 300 倍的趋势增长，通过增加模型参数和训练数据仍是短期内的发展方向。

2022 年 11 月，OpenAI 推出建立在 GPT-3 系列大型语言模型之上的 ChatGPT，并使用监督和强化学习技术进行微调。ChatGPT 在技术路径上采用“大数据+大算力+强算法=大模型”的战略，又在“基础大模型+指令微调”方向探索出新道路，基础大模型类似人类的大脑，通过指令微调进行交互训练，结合实现近似人类的语言智能。ChatGPT 的问世不仅是新一代聊天机器人的突破，还将为人工智能和整个信息产业带来一场革命。

图 12：超大规模模型参数和数据规模变化



资料来源：中国信通院，中国银河证券研究院

2、数据

在算法模型发展的同时，对于数据规模和质量的要求也在不断提高。以 GPT 的发展历程来看，用以训练模型的数据集的广度和深度都在不断加强，使得模型的回答具有更高的准确性和质量，实现模型的不断优化。

GPT 使用 BooksCorpus 数据集来训练语言模型。BooksCorpus 有大约 7000 本未出版的书籍，有助于在未见过的数据上训练语言模型。另外，这个语料库有大量的连续文本，有助于模型大范围地学习依赖关系。GPT-2 使用的训练数据集名为 WebText，具有来自 800 多万份文件的文本数据，总规模为 40GB，与用于训练 GPT-1 模型的图书语料库数据集相比是巨大的。GPT-3 是在五个不同的语料库中混合训练的，每个语料库都有一定的权重。其中高质量的数据集被更频繁地取样，并且不止被训练过一个 epoch。使用的五个数据集是 Common Crawl, WebText2, Books1, Books2 和 Wikipedia。用于训练 ChatGPT 的具体数据集没有公开披露，但仍然是几个大型语料库的组合，并且数据规模比 GPT-3 进一步增大。

表 3：GPT 训练数据规模持续增大

模型	参数量	数据规模	Tokens	数据来源
GPT	1.17 亿	5GB	117 百万	网页、书籍和文章，BooksCorpus 数据集
GPT-2	15 亿	40GB	15 亿	网页、书籍和文章，WebText 数据集
GPT-3	1750 亿	45TB	1750 亿	Common Crawl，WebText2，Books1，Books2,Wikipedia 数据集
ChatGPT	1750 亿	>45TB	7740 亿	多样化的文本来源，多个大型语料库的组合

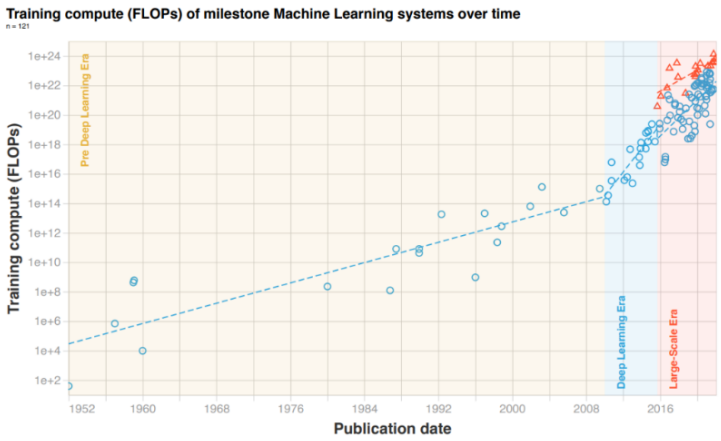
资料来源：ChatGPT，中国银河证券研究院

3、算力

自从进入互联网时代，人类所能获取和利用的数据呈现爆发式地增长，各行业、各场景的

海量数据为人工智能的自主学习和模型训练提供了数据基础。而自人工智能的概念兴起，算法模型一直在不断优化，从决策树到神经网络，从机器学习到深度学习，并且已在不同的领域中得到应用。算力是基于芯片的人工智能发展的硬件基础和平台，随着海量数据的产生和算法模型的不断优化和发展，算力的发展成为了人工智能系统快速发展的核心要素。从 1956-2020 年，计算机处理能力的 FLOPS 增加了一万亿倍。

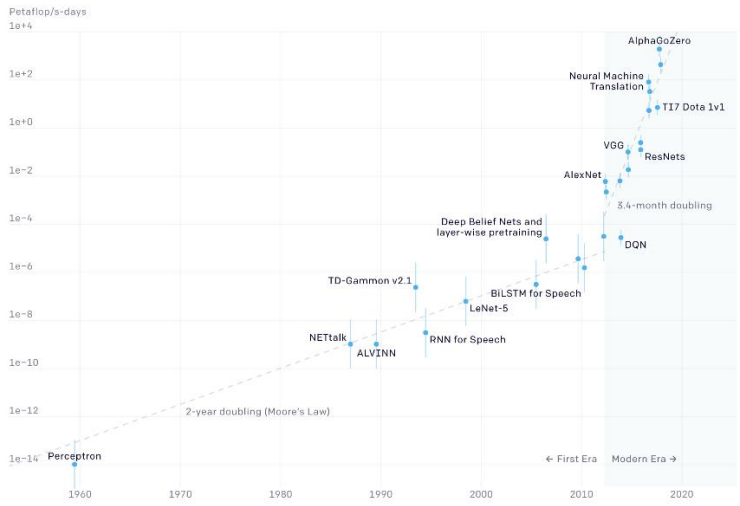
图 13：1956-2015 年算力实现万亿倍增长



资料来源：Experts-Exchange，中国银河证券研究院

近几年，大量复杂的数据的收集和处理都需要硬件能力的相应增长，以应对人工智能发展的需求。基本上，计算能力是计算机以速度和准确性执行某种任务的能力。正如 OpenAI 的研究表明，训练最大的人工智能模型所需的计算能力，自 2012 年以来平均以每 3.4 个月翻一倍的速度增长。而在 2012 年之前的情况并非如此，当时计算能力平均以 2 年的速度翻倍。这意味着，今天使用的资源正以比以前快七倍的速度翻倍。从另一个角度而言，在线性尺度上，计算用量在 2019 年之前就增加了 30 万倍，表明对人工智能特定硬件的需求呈指数级增长。

图 14：AlexNet 到 AlphaGo Zero：计算量增加 300,000 倍



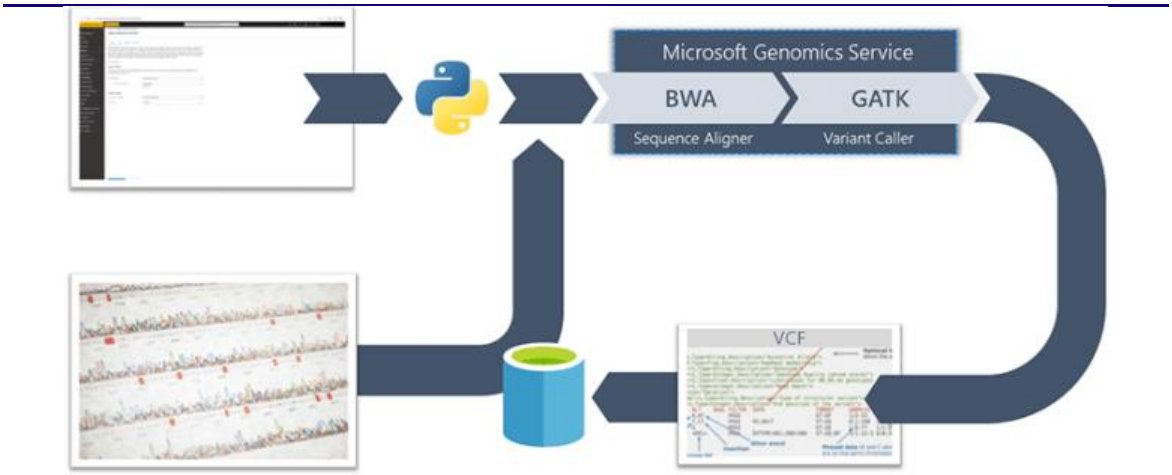
资料来源：OpenAI（一个 petaflop/s-day 包括在一天内每秒执行 1015 次神经网络操作），中国银河证券研究院

（四）互联网行业巨头积极布局，AI 竞赛压力不减

1、微软——投资 OpenAI，探索 AI 在在多场景落地

14 年起推动 AI 领域布局，逐步探索 AI 商业模式落地。作为互联网行业的领先者，微软过去专注于继续开发 Windows 和 Office 应用程序。2014 年，随着首席执行官 Satya Nadalla 的任命，微软开始向人工智能战略转变，推动微软在人工智能创新方面的发展。2016 年，微软成立了人工智能实验室，致力于推广和开发基于人工智能的应用程序。2017 年，微软宣布收购于以深度学习为研究重点的初创公司 Maluuba，并将人工智能的运用延伸到空中。同年，微软和亚马逊宣布建立合作伙伴关系，意味着微软人工智能开发的工具和服务，如 Cortana，Office 365 将与 Alexa 等亚马逊服务交互。2018 年，微软又相继收购多个 AI 公司，探索深度学习的商业化模式。

图 15：微软 AI 的基因组学小组在癌症治疗中的作用



资料来源：Algorithm-Xlab，中国银河证券研究院

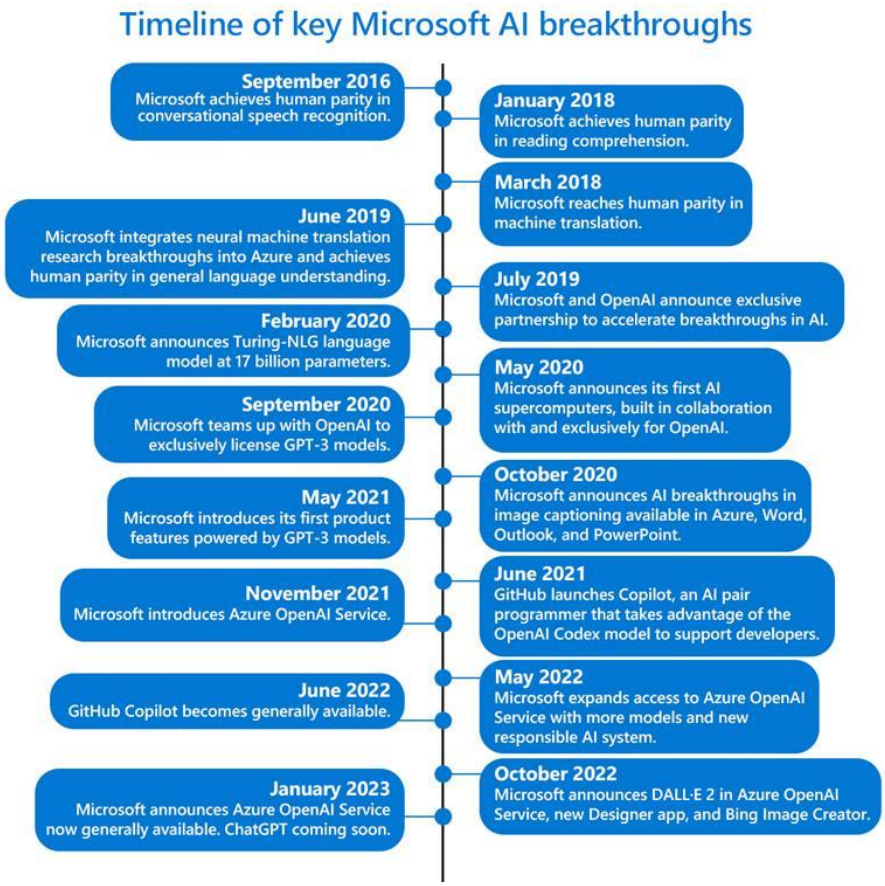
多次投资 OpenAI，在人工智能领域探索更进一步。

2019 年，在微软于首次向 OpenAI 注资后，两家公司开始在微软的 Azure 云计算服务上合作开发 AI 超级计算技术，同时，OpenAI 也逐步将其云服务从谷歌云迁移到 Azure。有了微软的算力支持，OpenAI 在 2020 年推出了突破性的成果 GPT-3。同样在 2020 年，微软买断了 GPT-3 基础技术的授权，在 Office、搜索引擎 Bing 和设计应用 Microsoft design 等产品中使用 GPT-3，以优化现有产品。

2021 年，微软再次投资，双方合作正式进入第二阶段。一方面，作为 OpenAI 的云服务商，微软在 Azure 中集中部署 GPT、DALLE、Codex 等 OpenAI 开发的各类工具。这也形成了 OpenAI 最早的收入来源，即通过 Azure 向企业提供付费 API 和 AI 工具。同时，在获得 OpenAI 新技术商业化许可的情况下，微软开始将 OpenAI 工具与自己的产品深度整合，并推出相应的产品。2021 年 6 月，微软与 OpenAI 和 GitHub 合作，推出了基于 Codex 的 AI 代码补充工具 GitHub Copilot，于次年 6 月上线。2023 年，微软向 OpenAI 追加投资数十亿美元，彻底拉开

了人工智能军备竞赛的帷幕，同时微软将 ChatGPT 整合到其搜索引擎中，标志着 OpenAI 新技术的商业化进入新阶段。

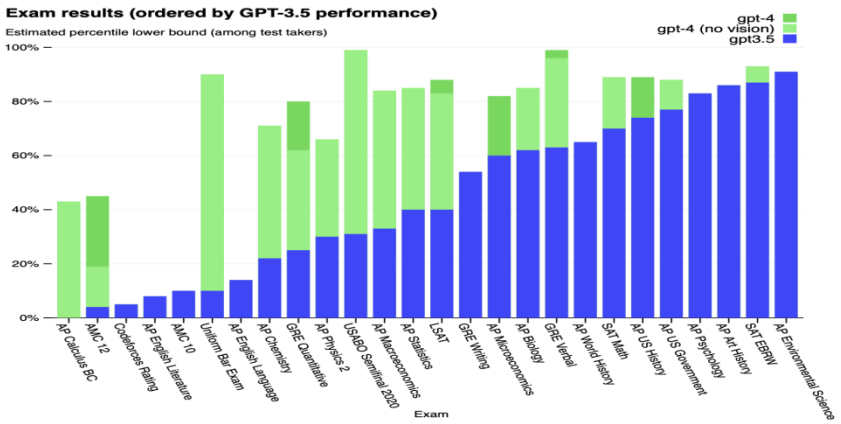
图 16：微软人工智能方面突破的时间轴



资料来源：微软官网，中国银河证券研究院

在 2023 年 3 月，OpenAI 又推出了 ChatGPT 的升级版——GPT-4，迭代速度极快。其包含的重大升级是支持图像和文本的输入，并且在 GPT-3 原来欠缺的专业和学术能力上得到重大突破，它通过了美国律师法律考试，并且打败了 90%的应试者。在各种类型考试中，GPT-4 的表现都优于 GPT-3。

图 17：GPT-4 与 GPT-3 在各类考试中的结果对比



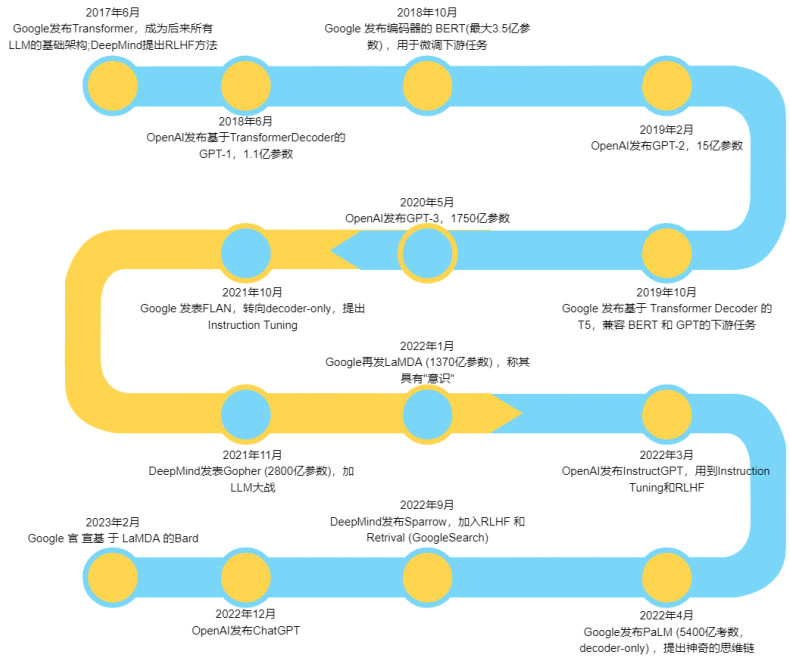
资料来源：OpenAI，中国银河证券研究院

2、谷歌——引领人工智能驱动商业化创新

谷歌是人工智能发展中最重要的公司之一。2010 年谷歌推出其第一个人工智能驱动的搜索引擎算法，称为 Google Instant。2012 年，谷歌推出知识图谱，首次使用人工智能来理解不同实体之间的关系。2015 年，谷歌推出了 TensorFlow，用于机器学习的开源软件库。2016 年，谷歌 DeepMind 的 AlphaGo 程序在围棋比赛中击败了世界冠军李世石。2017 年，谷歌推出了谷歌助理，一个可用于智能手机和智能家居设备的对话式人工智能助理。

自 2017 年，谷歌发布 Transformer 以来，NLP 领域的技术发展得到了质的飞跃，基于谷歌的成果，OpenAI 在 2018 年发布了 GPT 生成式预训练模型，也就是基于 Transformer Decoder 的 GPT-1，带有 1.1 亿参数，通过大规模、无监督的预训练+有监督的微调，在大型数据集上进行训练而建立的模型。与此同时，在 2018 年 10 月，谷歌推出了具有开创性的 BERT 模型，具有 3.4 亿个参数，比 GPT 大四倍并几乎在所有性能方面都超越 GPT。

图 18：谷歌 LLM 领域的发展时间轴

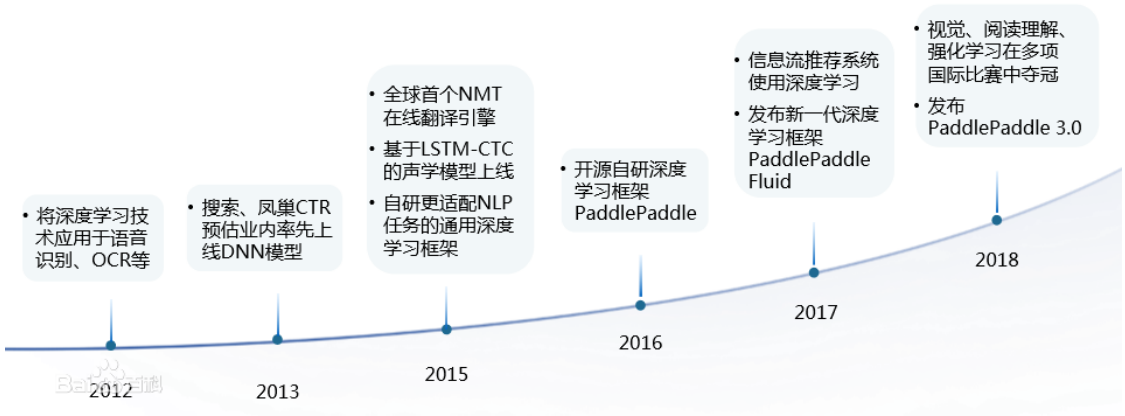


资料来源：智源社区，中国银河证券研究院

3、百度——All in AI，十年布局长跑

All in AI，十年布局长跑。百度在 AI 领域的布局早在 2010 年前就开始了。百度早在 2010 年代初就开始投资于人工智能技术。2014 年，百度成立了深度学习研究院（IDL），专注于开发深度学习算法和其他 AI 技术。2015 年，百度的语音识别软件实现了 5.5%的最低单词错误率（WER）记录。这是语音识别技术发展的一个重要里程碑，确立了百度在该领域的领先地位。在 2016 年 9 月的百度世界大会上，整合了视觉、语音、自然语言处理、知识图谱、深度学习等技术的百度大脑正式对外开放。2017 年，百度推出了阿波罗自动驾驶汽车平台。该平台为开发者提供一系列工具和资源，以建立自动驾驶系统。2018 年，百度推出了名为百度健康的医疗部门。该部门专注于使用人工智能来改善医疗诊断、药物开发和医疗保健的其他方面。2018 年，百度发布了其 AI 芯片“昆仑芯”，该芯片旨在用于人工智能应用，如自动驾驶和语音识别。

图 19：百度深度学习发展历程



资料来源：百度百科，中国银河证券研究院

百度在 AGCI 中的全栈布局：算法、算力、数据、应用。百度官方宣布：文心一言云服务于 2022 年 3 月 27 举行新品发布会。官方展示了文心一言在文学创作、商业文案创作、数理推算、中文理解、多模态生成五个使用场景中的综合能力。在文心一言的背后，是经过四年迭代的文心大模型。文心 ERNIE 自 2019 年诞生至今，在语言理解、文本生成、跨模态语义理解等领域取得多项技术突破，在公开权威语义评测中斩获了十余项世界冠军。文心模型的训练是基于百度飞桨的框架，在飞桨技术不断迭代的基础上，文新大模型一次性发布了 11 个大模型，涵盖了基础大模型、任务大模型和行业大模型三个层次的体系，充分满足了行业的应用需求。例如用于语言生成的 ERNIE 3.0 Titan，用于文本和图像生成的 ERNIE-ViLG 模型。目前，百度飞桨凝聚了 265 万开发者、服务了 10 万家企业、创建了超过 34 万个模型。

图 20：文心发展历程

资料来源：百度飞桨官网，中国银河证券研究院

百度自行研发的 AI 芯片，单卡算力达到 128TFLOPS。昆仑芯是基于百度在人工智能领域多年的产业实践，自主研发的一款人工智能通用处理器芯片。新发布的 R200 人工智能加速卡基于第二代昆仑芯，采用领先的 7nm 工艺，基于先进的芯片架构，专为深度学习和机器学习算法的云端和边缘计算设计。与上一代产品相比，R200 全面提升了计算机视觉、自然语言处理、大规模语音识别、大规模推荐等应用的人工智能负载的运行效率。

表 4：两代产品参数对比

型号	K100 加速卡	R200 加速卡

精度	INT4/8/16 XFP16/32	INT8/16/32 FP16/32
算力	INT8：128 TOPS FP16：32 TOPs FP32：8 TOPS	INT8: 256TOPS FP16: 128TOPS FP32: 32TOPS
显存	8GB	16GB
访存宽带	256 GB/s	512 GB/s
系统互联	PCI-E Gen4 x 8 ， 兼容 3.0/2.0/1.0	PCI-E Gen4 x 8 ， 兼容 3.0/2.0/1.0
功能	75W	150W

资料来源：百度智能云

在数据层面，百度基于其搜索引擎业务，积累了大量的真实用户数据。这些大规模的数据，使文心一言形成自身优势，为文心大模型的训练提供数据基础。产品中文心一格和文心百中已成功落地。文心一格是一个 AI 艺术和创意辅助平台，文心百中是大模型驱动的产业级搜索系统。

图 21：文心大模型全景图

资料来源：百度飞桨公众号，中国银河证券研究院

AIGC 的全球巨头争夺战已经开始，深耕 AI 和搜索领域多年的百度正站在一个新的历史舞台上，将于 2019 年 3 月推出的文心一言模型，是中国科技力量参与全球 AIGC 竞赛的主要代表。

二、英伟达举办 GTC2023，关注高性能计算相关领域壁垒

Navida 召开 GTC 发布会，展示算力芯片在多领域的突破进展。2023 年 3 月 21 日，英伟达召开 GTC，CEO 黄仁勋进行了主题演讲，展示英伟达算力芯片咋 AI 应用、加速卡领域取得进展，目前已成为自然科学、化学制药、视觉解析、数据处理、 机器学习和大模型领域成为不可或缺的一环。

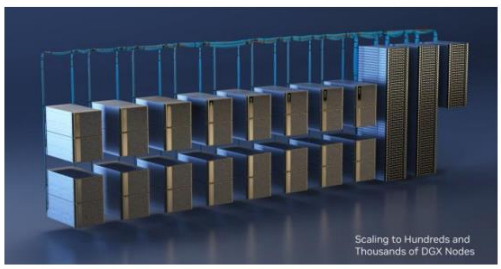
图 22：英伟达 GTC2023 会议

资料来源：GTC2023，中国银河证券研究院

AI 产业迎来“iPhone”时刻,英伟达 DGX 计算机已成 AI 核心处理器。目前英伟达已向 OpenAI 交付首台 DGX AI 超级计算机，用于加速深度学习、人工智能应用,《财富》100 强企业中已有一半以上企业开始使用 DGX，例如：BMW 应用 DGX 被用于加速 BMW 汽车自动驾驶系统的开发和训练；Tencent 应用 DGX 被用于加速腾讯云的人工智能服务的开发和运营；美国国家航空航天局利用 DGX 被用于加速 NASA 进行气象和环境数据的分析和预测。

从参数上来看，DGX 具备满足高性能计算和 AI 学习的需求。GPU 采用 8 片英伟达 A100 Tensor Core GPU，共有 6912 个 CUDA 核心和 432 个 Tensor Core，单精度计算性能为 320TFlops.CPU 采用两颗英特尔 Xeon Platinum 8280L 处理器，共有 56 个核心；每个 DGX 系统配备 1.5TB 的 DDR4 内存；每个 DGX 系统配备 15TB 的 NVMe 存储器，同时支持 100Gb Ethernet 和 Infiniband HDR 网络。DGX 具有强大的计算性能、高效的数据传输速度、大容量的存储空间和稳定的供电系统，能够满足各种深度学习和人工智能应用的需求

图 23：上万台 DGX 连接组成 AI 超级计算机



资料来源：GTC2023，中国银河证券研究院

图 24： DGX A100 系统与 AI 数据中心参数比较

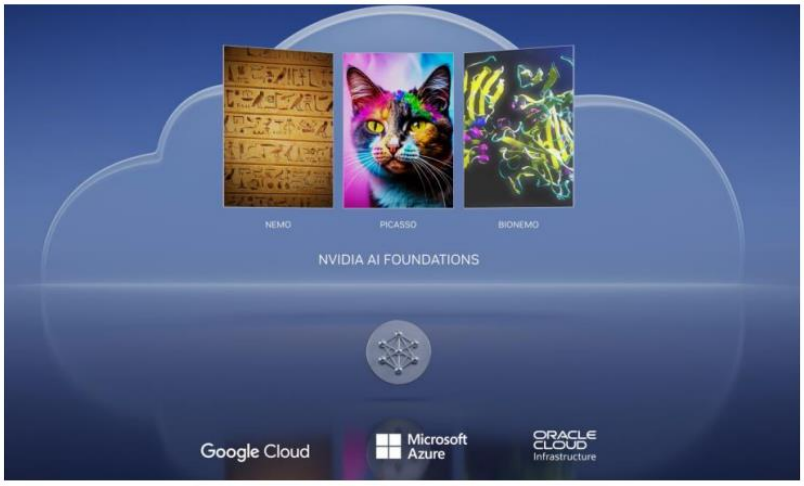
	传统的AI数据中心	DGX A100	VS 传统的AI数据中心
AI训练	50个DGX-1	5个DGX A100系统	
AI推理	600个CPU系统		
资金开销	1100万美元	100万美元	1/10
机架数量	25个机架	1个机架	1/25
功率	630kW	28kW	1/20

资料来源：智东西，中国银河证券研究院

英伟达推出 AI Foundations 云服务，从 NEMO、PICASSO、BIONEMO 三方面渗透 AI

场景。AI Foundations 一站式云服务，从模型的构建到生成应用上线，，协助客户快速构建、优化和运营大模型，把制造大模型的能力传递到每一个用户。

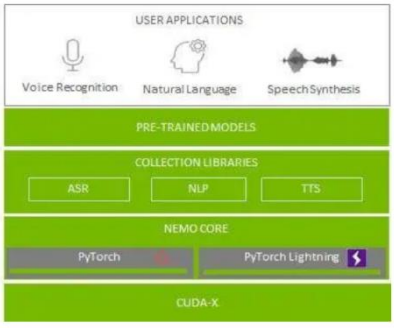
图 25：AI Foundations 一站式云服务



资料来源：GTC2023，，中国银河证券研究院

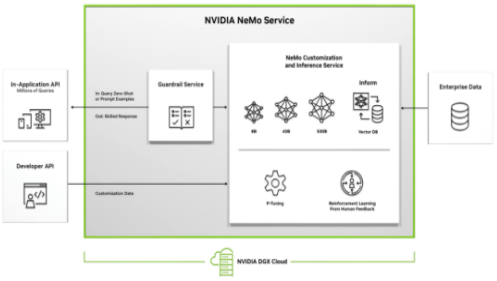
NVIDIA NeMo 是一个基于 PyTorch 的开源工具包，用于自然语言文本的生成式模型。提供 80 亿、430 亿、5300 亿参数的 GPT 模型，客户也可以引入自己想要的模型。Nemo 会定期更新额外的训练数据，可以帮助企业为客服、企业搜索、文档处理、市场分析等场景定制生产生成式 AI 模型。

图 26：NeMo 的应用栈



资料来源：英伟达，中国银河证券研究院

图 27：NEMO 服务流程



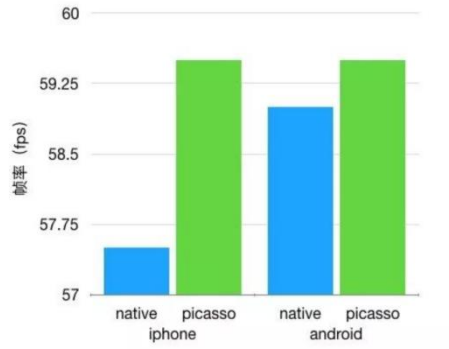
资料来源：英伟达，中国银河证券研究院

PICASSO (Parallel-n-Core Architecture Simulator for Scalable Oltp) 是一个用于模拟大规模多核处理器架构的开源模拟器，用于训练能够生成图像、视频和 3D 素材的模型。NVIDIA 与 Adobe 宣布扩展双方的长期研究和开发合作关系，共同推动下一代生成式 AI 模型，为加快优秀创作者和营销人员的工作流程，其中一些模型将采取联合开发的方式，并 NVIDIA Picasso 进入市场。同时，NVIDIA 正与 Getty Images 联合训练负责任授权的生成式文本转图像以及文本转视频基础模型，这些模型将使用简单的文本提示创建图像和视频，并将在 Getty Images 完全授权的资产上进行训练。

图 28：Picasso 动态化原理

资料来源：英伟达，中国银河证券研究院

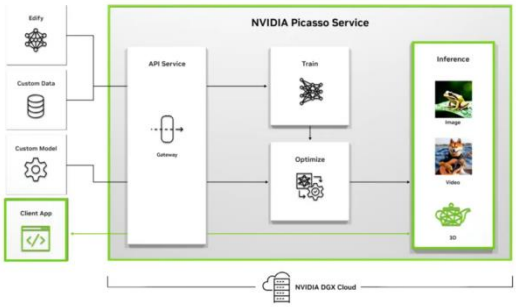
图 29：Picasso 高性能渲染



资料来源：英伟达，中国银河证券研究院

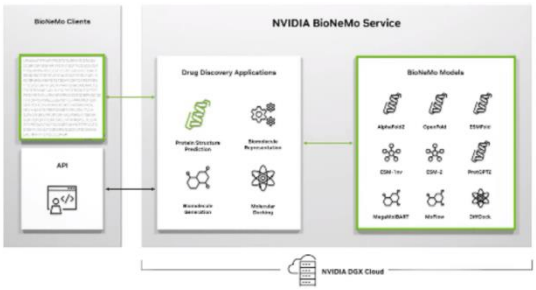
BioNeMo 服务提供用于化学和生物学的 LLM。NVIDIA BioNeMo 框架用于训练和部署超算规模的大型生物分子语言模型，帮助科学家更好地了解疾病，并为患者找到治疗方法。该大型语言模型（LLM）框架将支持化学、蛋白质、DNA 和 RNA 数据格式。

图 30：BIONEMO 提供多种生物制药领域模型



资料来源：英伟达，中国银河证券研究院

图 31：BIONEMO 支持云端运行



资料来源：英伟达，中国银河证券研究院

发布 H100 NVL 服务器，相比 A100 DGX 提供 10 倍的计算速度。GTC2023 同时发布 H100 NVLINK，这款 H100 GPU 启用了基本完全的 94GB HBM 显存堆栈。最大区别在于，双 GPU 结构，顶部使用 3 个 NV Link 连接器进行互联，因此可以提供多达 188GB 显存，显存带宽也不止翻倍，每个 GPU 带宽提供 3.9TB/s，而 H100 SXM 为 3.35TB/s，H100 PCIe 为 2TB/s。**H100 NVL 综合性能可以达到 H100 SXM 的两倍。**

图 32： Nvidia 不同显卡类型规格对比

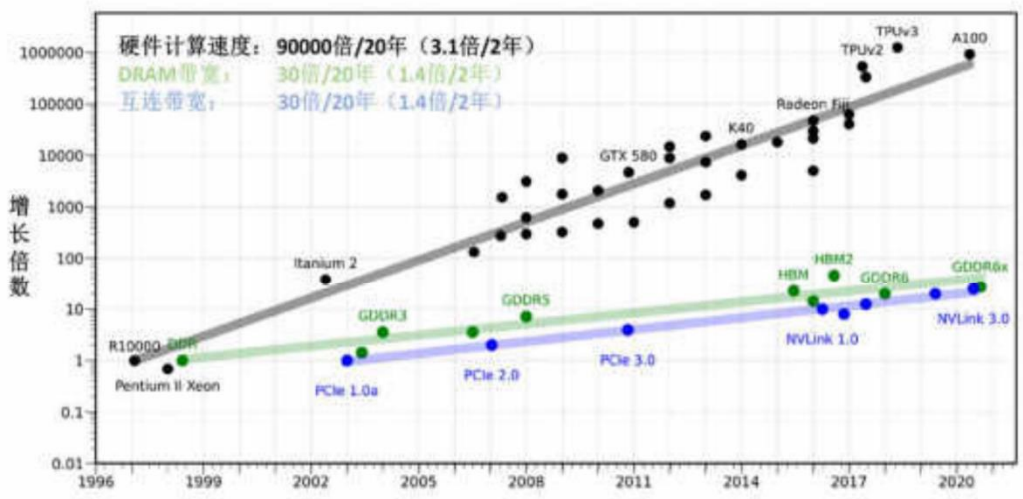
VideoCardz.com	NVIDIA H100	NVIDIA A100	NVIDIA Tesla V100	NVIDIA Tesla P100
Picture				
GPU	GH100	GA100	GV100	GP100
Transistors	80B	54.2B	21.1B	15.3B
Die Size	814 mm²	828 mm²	815 mm²	610 mm²
Architecture	Hopper	Ampere	Volta	Pascal
Fabrication Node	TSMC N4	TSMC N7	12nm FFN	16nm FinFET+
GPU Clusters	132/114*	108	80	56
CUDA Cores	16896/14592*	6912	5120	3584
L2 Cache	50MB	40MB	6MB	4MB
Tensor Cores	528/456*	432	320	-
Memory Bus	5120-bit	5120-bit	4096-bit	4096-bit
Memory Size	80 GB HBM3/HBM2e*	40/80GB HBM2e	16/32 HBM2	16GB HBM2
TDP	700W/350W*	250W/300W/400W	250W/300W/450W	250W/300W
Interface	SXM5/*PCIe Gen5	SXM4/PCIe Gen4	SXM2/PCIe Gen3	SXM/PCIe Gen3
Launch Year	2022	2020	2017	2016

资料来源：英伟达，中国银河证券研究院

（二）大算力场景下，多项技术瓶颈期待突破

大算力背景下，存算性能呈现剪刀差，存储器件性能远弱于算力性能提升。随着 AI 算力需求的不断提升，传统存储器件也到达了尺寸的极限。依靠先进制程工艺不断缩小器件面积、同时提升算力的方式似乎已经走入死路。我们突破 AI 算力困境的方式，有着两条清晰的路线：架构创新与存储器件创新。“存”“算”之间性能失配，从而导致了访存和成本优化，带宽低、时延长、功耗高等问题，即通常所说的“存储墙”和“功耗墙访存愈密集，“墙”的问题愈严重算力提升愈困难。随着以人工智能为代表的访存密集型应用快速崛起访存时延和功耗开销无法忽视，计算架构的变革显得尤为迫切。

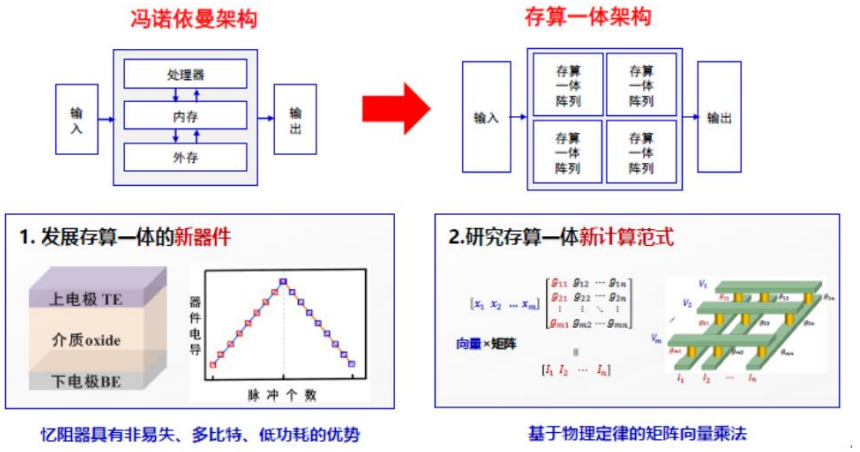
图 33：存储计算“剪刀差”



资料来源：《AI and Memory Wall》，中国银河证券研究院

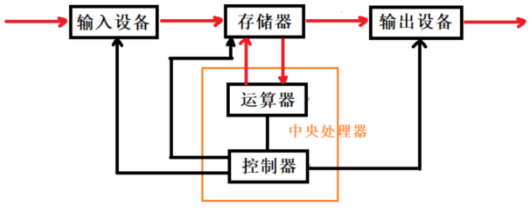
冯诺依曼架构，导致数据传输的 90%功率消耗都在数据传输上，99%的时间都消耗在存储器读写过程中，导致“存储墙”和“功耗墙”问题。冯诺依曼架构的芯片在工作时，计算单元要先从内存中读取数据，计算完成后再存回内存，才能最终输出。在过去，存储器与处理器的发展严重失衡，自上世纪八十年代以来，存储器读取速率的提升远远跟不上处理器性能的增长。这导致了计算畸形的漏斗结构：无论处理器所在的漏斗“入口”一端处理了多少数据，也只能通过存储器狭窄的“出口”输出，严重影响了数据处理的效率。

图 34：基于忆阻器的存算一体技术



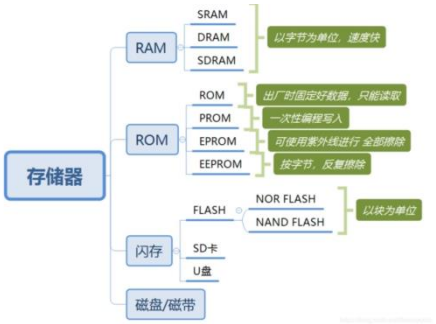
资料来源：中国银河证券研究院

图 35：冯诺依曼计算架构



资料来源：《存算一体白皮书》，中国银河证券研究院

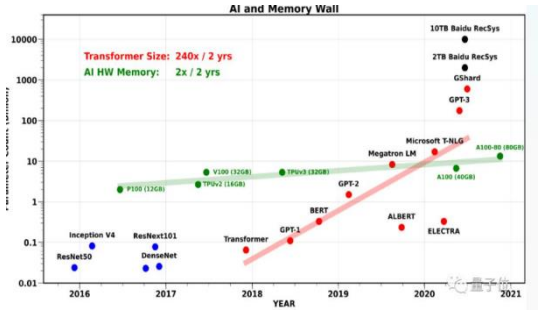
图 36：存储器类型



资料来源：中国银河证券研究院

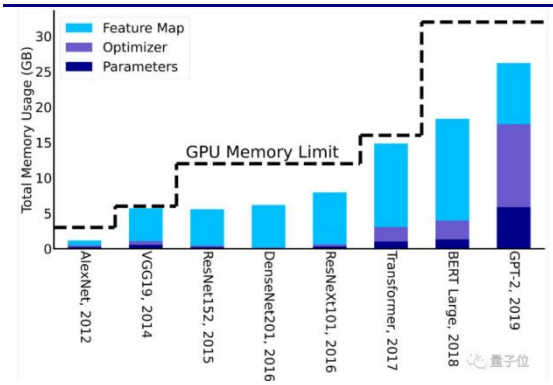
AI 训练未来的瓶颈不是算力，而是 GPU 的“内存墙”。无论是芯片内部、芯片间，还是 AI 加速器之间的通信，都已成为 AI 训练的瓶颈。其中，Transformer 模型中的参数数量（红色）呈现出 2 年 240 倍的超指数增长，而单个 GPU 内存（绿色）仅以每 2 年 2 倍的速度扩大。尽管在日常 GPU 使用中，对“内存墙”的存在并不敏感，但是 AI 模型的内存需求，通常是参数数量的几倍。因为训练需要存储中间激活，通常会比参数（不含嵌入）数量增加 3-4 倍的内存。于是，AI 训练不可避免地撞上了内存容量以及内存传输带宽的墙。

图 37：AI 模型和 GPU 内存增长剪刀差



资料来源：量子位公众号，中国银河证券研究院

图 38：非 AI 训练传输容量和速度没有触摸到内存墙



资料来源：量子位公众号，中国银河证券研究院

（三）存算一体化趋势确定，HBM 与 Chiplet 实现降本增效

全球半导体厂商已提出多种解决方案，存内计算电路可基于 SRAM 和 NOR Flash 实现。AI 对数据的访问和不断调取需要数据需要在存储单元和计算单元之间频繁移动，访存带宽和功耗成为算法的重要瓶颈之一。存算一体将存储单元与计算单元直接结合在一起，绕过数据在存储和计算之间的搬运环节。当前 NOR Flash、SRAM 等传统器件相对成熟可率先开展存内计算产品化落地推动，从方案落地情况来看，英特尔选择基于 SRAM 的可配置存储器，三星选择在 DRAM 的 DRISA 架构上进行存算一体解决方案。

表 5：存内计算器件对比分析

器件	SRAM	NOR Flash	RRAM	MRAM	PCM
易失特性	易失	非易失	非易失	非易失	非易失
多值存储	否	是	是	否	是
现有工艺节点	5nm	28nm	28nm	16nm	28nm
理论工艺极限	2nm	14nm	5nm	5nm	5nm
单比特存储面积 (F²/bit')	~300	~7.5	20~40	~30	~24
读写次数	无限	10^6	10^8	~10^15	10^8
应用场景	云侧和边侧的推理和训练	边侧和端侧的推理	云侧、边侧和端侧的推理	云侧和边测的推理和运算	云侧、边侧和端侧的推理

资料来源：《存算一体白皮书》，中国银河证券研究院

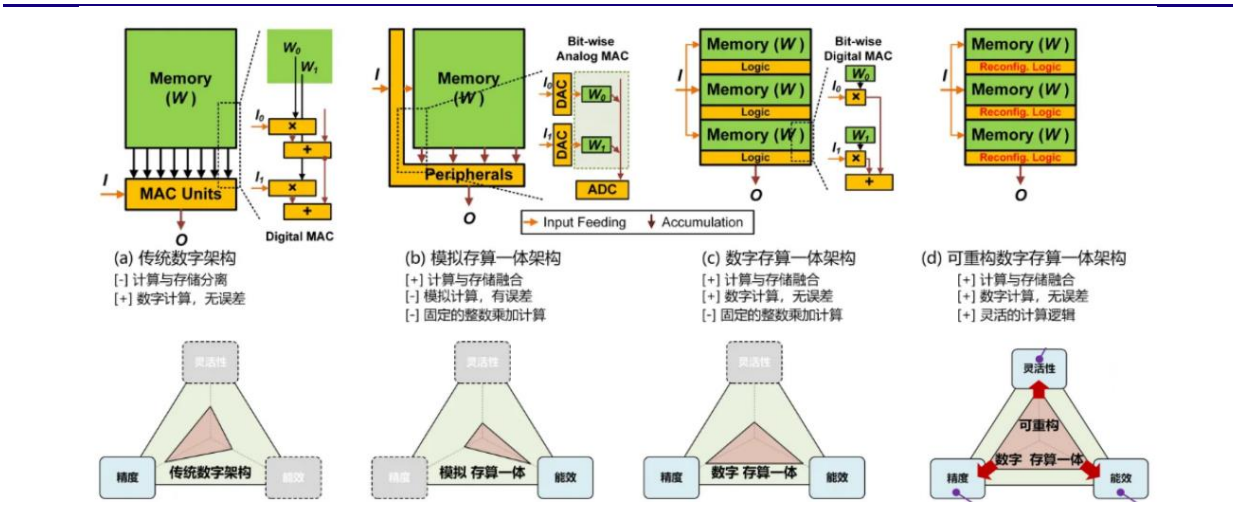
表 6：全球厂商的存算一体解决方案

厂商	存算一体解决方案
英特尔	基于 SRAM 的可配置存储器
三星	基于 DRAM 的 DRISA 架构
IBM	基于相变存储（PCM）的芯片设计方案
惠普	基于忆阻器实现逻辑存储融合
台积电	基于 ReRAM 的存算一体
知存科技	基于 NOR Flash 闪存的存算一体

资料来源：Intel，三星，IBM，HP，台积电，知存科技，中国银河证券研究院

存算一体架构可突破冯诺依曼瓶颈，提高 AI 芯片能效。存算一体架构消除了计算与存储的界限，直接在存储器内完成计算，被认为是突破冯诺依曼瓶颈的极具潜力的高能效 AI 芯片架构。目前主流的存算一体 AI 芯片基于模拟计算架构设计。模拟存算一体架构通常基于 SRAM 或非易失存储器，模型权重保持在存储器中，输入数据流入存储器内部基于电流或电压实现模拟乘加计算，并由外设电路对输出数据实现模数转换。由于模拟存算一体架构能够实现低功耗低位宽的整数乘加计算，非常适合边缘端 AI 场景。

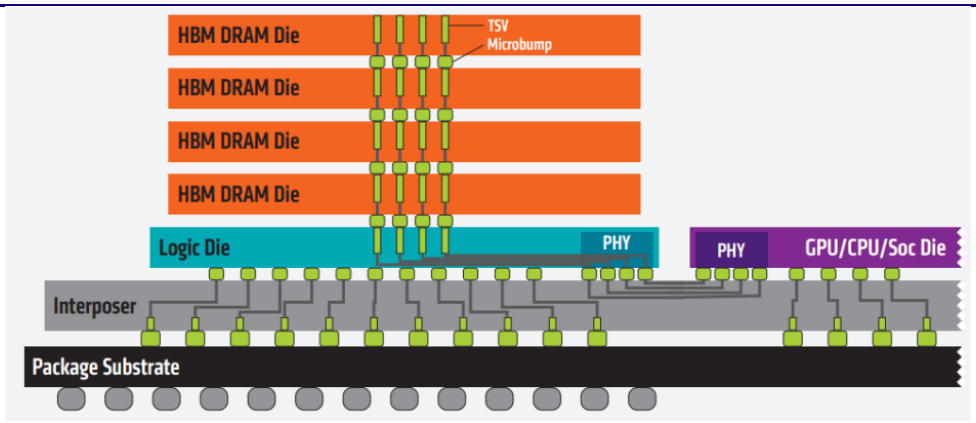
图 39：四种存算一体架构



资料来源：《ISSCC22 奇妙之旅》，中国银河证券研究院

HBM 的高带宽技术，从硬件上实现高速传输。高带宽存储器（HBM）可支持更高速率的带宽，基于 TSV 和芯片堆叠 技术的堆叠 DRAM 架构，可实现高于 256GBps 的突破性带宽，单颗粒的带宽远超过 DDR4 和 GDDR6。其中 DDR4 是 CPU 和硬件处理单元的常用外挂存储设备，8 颗 DDR4 颗粒带宽能够达到 25.6 GB/s，是 HBM 的 1/10，而 GDDR6 它单颗粒的带宽只有 64 GB/s，为 HBM 的 1/4。

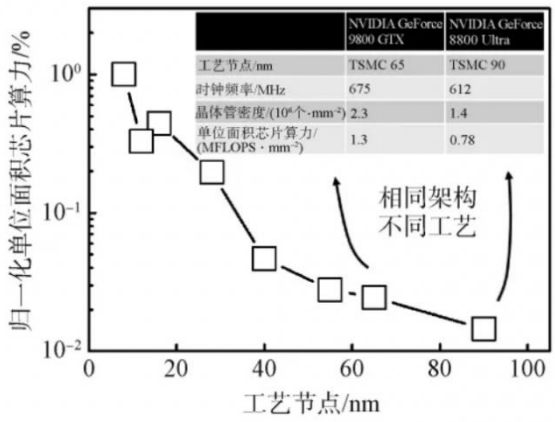
图 40：HBM 设计结构



资料来源：AMD，中国银河证券研究院

先进工艺是芯片算力提升的关键推动力，“后摩尔时代”先进封装不断发力。目前通过工艺提升芯片算力，主要有两种方式。1）先进制程：单位面积芯片算力会随着工艺节点的进步而提升，从 65nm 到 90nm 制程下的 GPU，先进工艺节点晶体管密度和工作频率均显著提高，从而带来芯片整体算力的提升。根据摩尔定律经验，集成电路上可以容纳的晶体管数目每 18 个月便会提升 1 倍，然而随着先进制程进入 3nm 时代，摩尔定律已经受到了物理极限和工艺成本的双重挑战。2）先进封装：先进封装可以优化连接方式、实现异构集成、提高芯片的功能密度，从而提升芯片算力，因而是超越摩尔定律方向中的重要赛道。21 世纪初，以 MEMS、TSV、FC 等为代表的先进封装技术引领封测行业发展，目前平面封装正在向 2.5D/3D chiplet 堆叠异构集成封装技术升级跃迁，为芯片算力提升带来了新思路。

图 41：英伟达 GPU 算力和工艺节点的关系



资料来源：《前瞻科技》，中国银河证券研究院

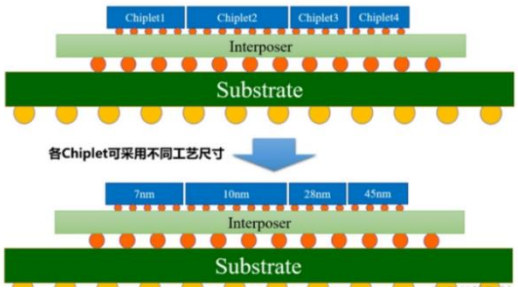
图 42：异构堆叠芯片图示

资料来源：世界半导体论坛，中国银河证券研究院

Chiplet 解决方案是底层基础，2.5D 和 3D 封装蓄势待发。Chiplet 技术是将大型单元芯片划分为多个相同或者不同的小芯片，这些小芯片可以使用相同或者不同的材质、工艺节点制造，再通过先进的集成技术封装在一起形成一个系统级芯片，降低成本的同时获得更高的集成度。目前寒武纪思元 370 系列产品就是在封装层面上，采用 Chiplet 技术，将两颗 370 芯片拼凑成算力更强、带宽更大的处理器模块。

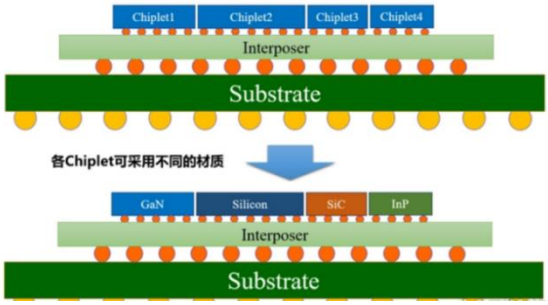
2.5D 封装技术是将芯片并排放置在中介层顶部，通过芯片的微凸块和中介层中的布线联系起来；3D 封装技术则无需中介层、芯片直接通过 TSV 直接进行高密度互连。通过 2.5D/3D 技术封装技术，可以在单位体积内集成更多的功能单元，并且这些功能单元之间互联很短，密度很高，因此性能可以得到很大的提升，算力水平也会提高。目前已有多家公司陆续布局 2.5D/3D 封装技术，封装领域将迎来又一次技术革命。

图 43：各 Chiplet 可采用不同工艺尺寸



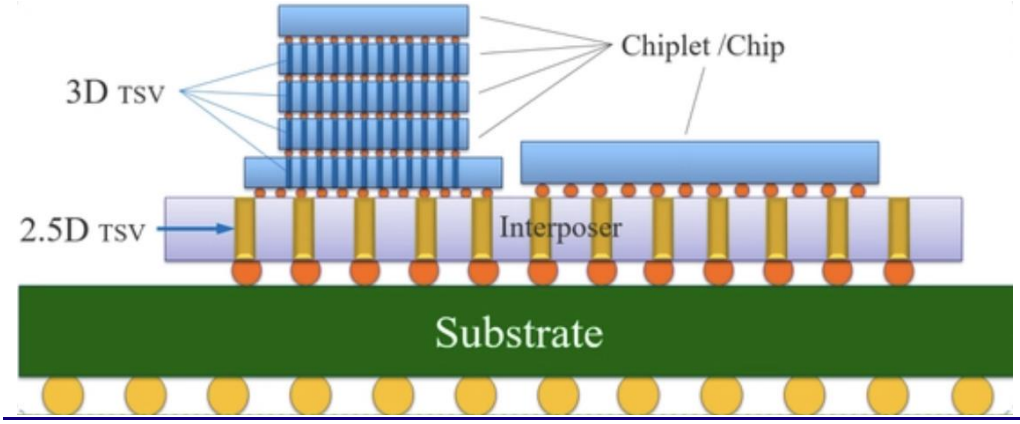
资料来源：芯榜科技，中国银河证券研究院

图 44：各 Chiplet 可采用不同材质



资料来源：龙芯中科官网，中国银河证券研究院

图 45：2.5D/3D 封装结构示意图

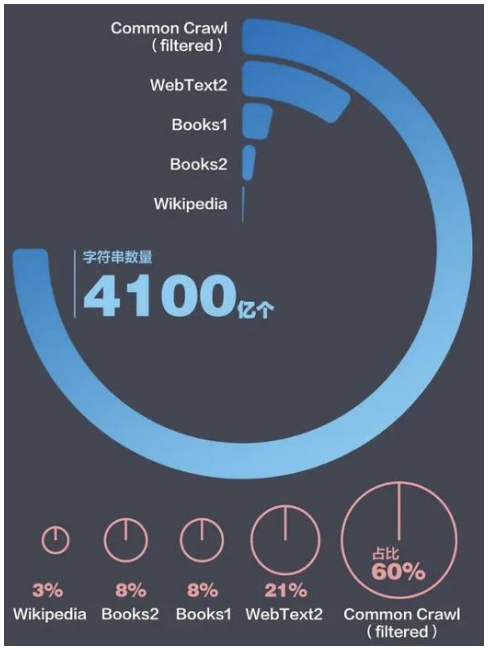


资料来源：芯榜科技，中国银河证券研究院

三、AI 商业落地曙光出现，ChatGPT 引爆大算力需求

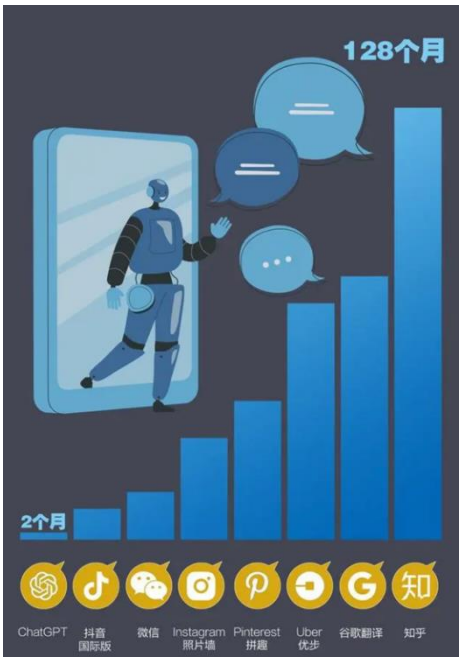
ChatGPT 是美国 OpenAI 公司开发的一款可实现精确问答的聊天机器人。ChatGPT 是由 GPT（Generative Pretrained Transformer）技术驱动，使用海量语料库进行训练的语言生成器。与其他语言生成器相比，GPT 技术采取了预训练生成器的方式，能够更好的理解人类语言的描述和数据中的知识，自动生成匹配内容且自然流畅的语言，并具有实现翻译、撰写邮件等各类语言相关任务的能力，大大提高了用户体验。因此，在 2022 年 11 月 ChatGPT 推出后，迅速引爆市场，2 个月内月活跃用户数便达一亿，成为了历史上用户增长最快的消费应用。

图 46： ChatGPT 海量数据的来源占比



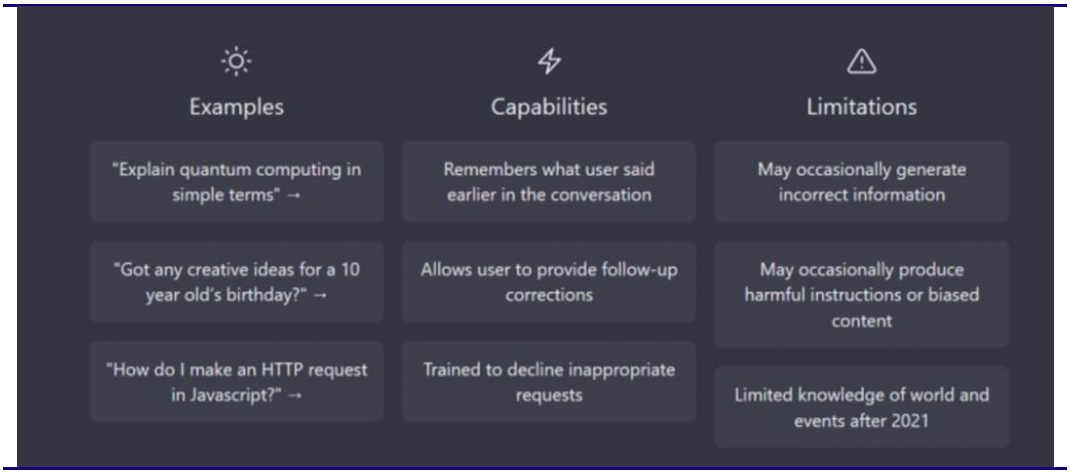
资料来源：Open AI 《Language Models are Few-Shot Learners》，瑞银集团，中国银河证券研究院

图 47：部分应用程序月活达到一亿所用时间



资料来源：瑞银集团，中国银河证券研究院

图 48：ChatGPT 的应用界面



资料来源：OpenAI，中国银河证券研究院

ChatGPT 参数量的提升代表了 AI 大模型的最新进展。AI 大模型（人工智能预训练大模型）指的是兼具“大规模（亿级参数）”和“预训练”两种功能属性的模型。从参数规模来看，AI 大模型的发展可以分为预训练模型、大规模预训练模型、超大规模预训练模型三个阶段。ChatGPT 的发展也反应了 AI 大模型的发展趋势，2018 年 OpenAI 发布的 ChatGPT 1.0 的模型参数为 1.17 亿，2019 年的第二代模型参数为 15 亿，ChatGPT 3.0 的参数相比于 ChatGPT2.0 增长了近百倍，达到了 1750 亿。

ChatGPT 的 AI 文本生成技术也是 AI 音视频、游戏等领域的底层技术，因此 ChatGPT3.0 的突破也将为 ChatGPT4.0 和 AIGC 领域提供更多的可能性，比如生成视频等。根据微软德国公司 CTO Andreas Braun 对 ChatGPT4.0 的预告，其参数量将为 3.0 的数倍，并拥有多模态模型。

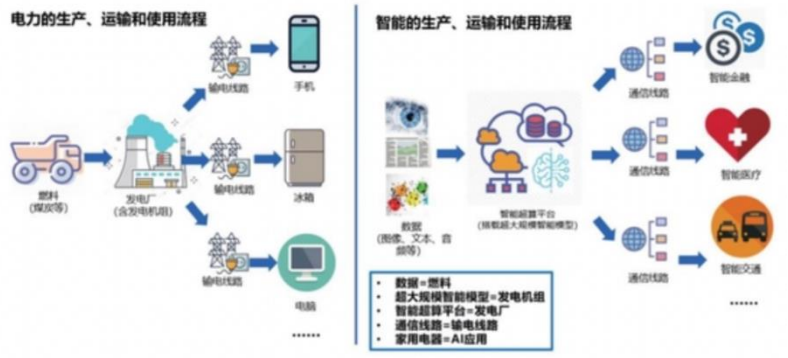
图 49：不同的语言模型训练所用的总算力、参数（Params）、训练数据量等（Token）

Model	Total train compute (PF-days)	Total train compute (flops)	Params (M)	Training tokens (billions)	Flops per param per token	Mult for bwd pass	Fwd-pass flops per active param per token	Frac of params active for each token
T5-Small	2.08E+00	1.80E+20	60	1,000	3	3	1	0.5
T5-Base	7.64E+00	6.60E+20	220	1,000	3	3	1	0.5
T5-Large	2.67E+01	2.31E+21	770	1,000	3	3	1	0.5
T5-3B	1.04E+02	9.00E+21	3,000	1,000	3	3	1	0.5
T5-11B	3.82E+02	3.30E+22	11,000	1,000	3	3	1	0.5
BERT-Base	1.89E+00	1.64E+20	109	250	6	3	2	1.0
BERT-Large	6.16E+00	5.33E+20	355	250	6	3	2	1.0
RoBERTa-Base	1.74E+01	1.50E+21	125	2,000	6	3	2	1.0
RoBERTa-Large	4.93E+01	4.26E+21	355	2,000	6	3	2	1.0
GPT-3 Small	2.60E+00	2.25E+20	125	300	6	3	2	1.0
GPT-3 Medium	7.42E+00	6.41E+20	356	300	6	3	2	1.0
GPT-3 Large	1.58E+01	1.37E+21	760	300	6	3	2	1.0
GPT-3 XL	2.75E+01	2.38E+21	1,320	300	6	3	2	1.0
GPT-3 2.7B	5.52E+01	4.77E+21	2,650	300	6	3	2	1.0
GPT-3 6.7B	1.39E+02	1.20E+22	6,660	300	6	3	2	1.0
GPT-3 13B	2.68E+02	2.31E+22	12,850	300	6	3	2	1.0
GPT-3 175B	3.64E+03	3.14E+23	174,600	300	6	3	2	1.0

资料来源：Language Model are Few-Shot Learners (2020)，OPENAI，中国银河证券研究院

AI 大模型突破传统 AI 适用性弱的局限，但是依旧面临商业化难的问题。传统的 AI 模型通常只针对性的针对一个或者一类任务，而 AI 大模型中大规模的参数量可以提升模型的表达能力，更好的建模海量训练数据中包含的通用知识，再通过“微调”使大模型在特定化的场景中依旧得到优越的表现。通过“预训练+微调”，AI 大模型已经具有强大的通用性，ChatGPT3.0 通过 prompt-tuning 免去微调步骤实现了更强的通用性。但是由于 AI 大模型的技术成本高昂并且决策过程难以解释，如何真正的商业化落地始终是 AI 产业中的难题。

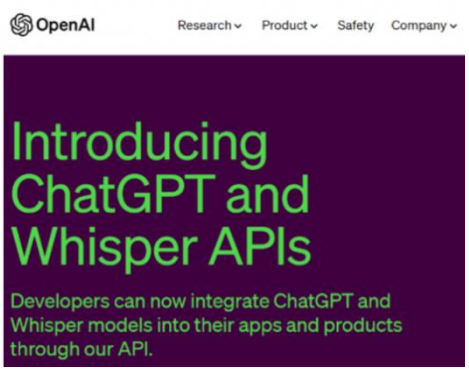
图 50：大模型具有较强的通用性，赋能 AI 到千行百业



资料来源：智源研究院，中国银河证券研究院

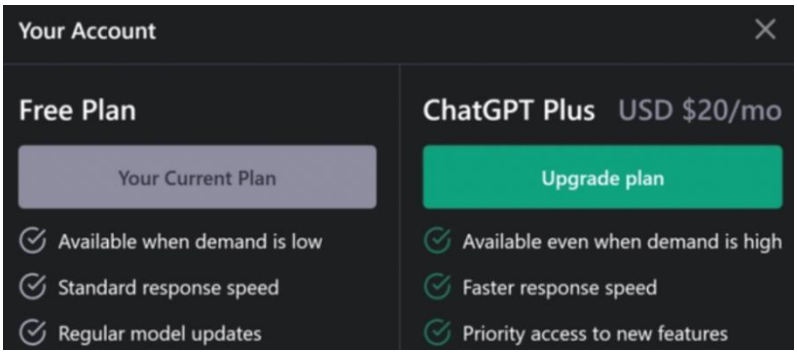
ChatGPT 率先在 C 端实现商业化，为 AIGC 产业落地带来曙光。2019 年 OpenAI 与微软合作，从非盈利性组织转为有限盈利公司，目前 ChatGPT 主要通过三种方式产生商业化收入。1）API 许可费：将 GPT-3 等模型开放给其他商业公司使用，根据用量收取费用。2）与微软深度合作：集成于微软云计算服务平台 Azure 和搜索引擎 Bing 上。3）订阅：推出付费订阅版 ChatGPT Plus，每月收费 20 美元。从 ChatGPT 的商业模式中，也可以看出生成式 AI 的 to C 端商业模式已经逐渐浮出水面，为 AIGC 产业实现商业化落地带来了新的希望。

图 51：ChatGPT 官宣开放 API 授权



资料来源：OpenAI，中国银河证券研究院

图 52：ChatGPT 推出付费订阅版 ChatGPT Plus

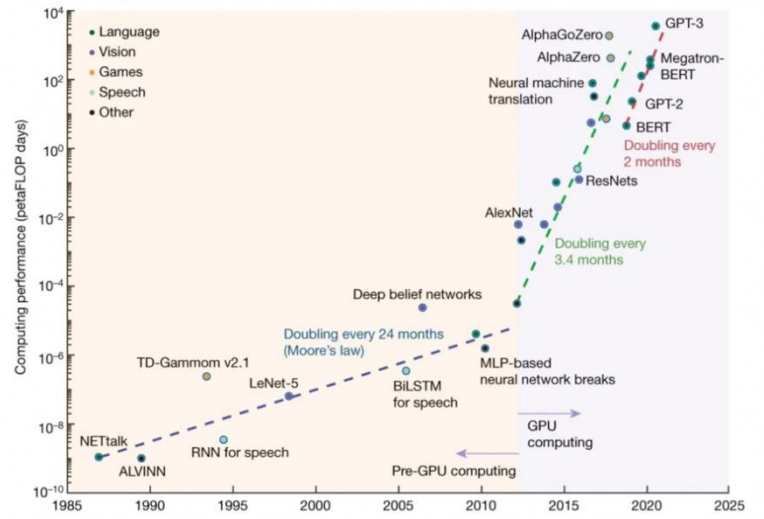


资料来源：公司招股说明书，中国银河证券研究院

（一）AI 芯片：算力水平是核心竞争力

强大的算力水平是AI 大模型必备的技术支撑。算力水平是数据处理能力强弱的决定性因素，AI 大模型的参数和语料库能够不断扩容离不开强大的算力支撑，根据英伟达的数据，ChatGPT 3.0 模型需要使用 1024 颗英伟达 A100 芯片训练长达一个月的时间。2012-2018 年，最大的 AI 训练算力消耗已增长 30 万倍，平均每 3 个多月便翻倍，速度远远超过摩尔定律。IDC 数据显示，2022 年中智能算力规模达到 268 百亿亿次/秒（EFLOPS），已经超过通用算力规模， AIGC 商业落地蓄势待发，未来对算力的需求更将超乎想象。

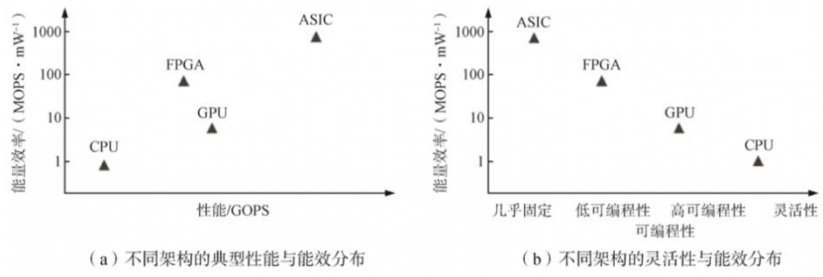
图 53：1985-2025 年间算力需求的增长



资料来源：Brain-inspired computing needs a master plan (2022)，中国银河证券研究院

GPU/ASIC/FPGA 三种计算架构并行。AI 芯片计算架构的好坏影响芯片能提供的算力水平，是决定芯片算力的本质因素。计算架构也需要在通用性和高效性之间进行平衡，目前 AI 芯片有 3 种主流计算架构，其中 GPU 计算架构在算力加速芯片中达到 90%。1）GPGPU：负责非图形相关程序的运算，具有高度可编程性，是最通用、最灵活的芯片，但是算力水平受限。2）ASIC：高定制化专用计算芯片，针对具体的应场景和算法，性能较高，但是通用性差 3）FPGA：基于现场可编程逻辑阵列的计算芯片，开发成本低、周期短，通用性和高效性介于 GPGPU 和 ASIC 之间。

图 54：GPU/ASIC/FPGA 三种计算架构特点



资料来源：电子发烧友，中国银河证券研究院

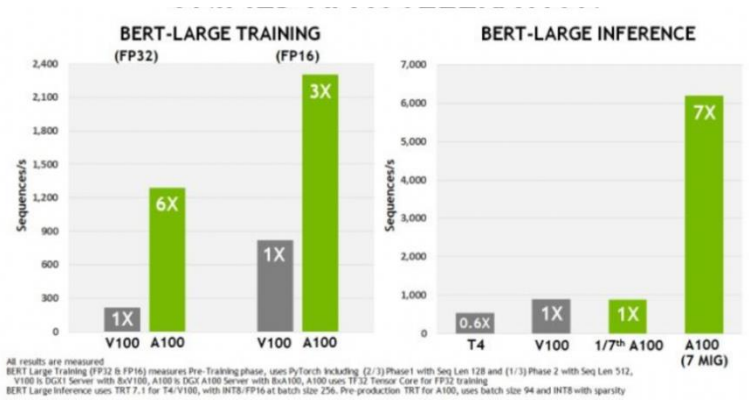
英伟达主导市场，国内厂商百花待放。目前算力芯片市场主要被欧美和日本厂商主导，其中英伟达是全球 GPU 领域的绝对龙头。英伟达 2020 年推出的 A100 芯片支持 FP16、FP32 和 FP64 浮点运算，峰值算力高达 624TOPS，预计在今年发布的 H100 芯片在 FP16、FP32 和 FP64 浮点计算方面将比 A100 快 3 倍，是当之无愧的 AI 芯片性能天花板。中国算力芯片领域起步较晚，但是在国家政策的大力扶持和企业持续的研发投入下，不少国内企业也在这方面取得了进展。

图 55：英伟达 A100 芯片规格参数

FP64 峰值性能	9.7 TF
FP64 Tensor Core 峰值性能	19.5 TF
FP32 峰值性能	19.5 TF
TF32 Tensor Core 峰值性能	156 TF | 312 TF*
BFLOAT16 Tensor Core 峰值性能	312 TF | 624 TF*
FP16 Tensor Core 峰值性能	312 TF | 624 TF*
INT8 Tensor Core 峰值性能	624 TOPS | 1,248 TOPS*

资料来源：RFID 信息，中国银河证券研究院

图 56：英伟达 A100 和英伟达其他芯片性能对比



资料来源：仪器小助手，中国银河证券研究院

寒武纪：中国 AI 芯片领导者。寒武纪成立于 2016 年，技术积累深厚，能提供云边端一体、软硬件协同、训练推理融合、具备统一生态的系列化智能芯片产品和平台化基础系统软件。近年来，公司持续加大研发投入，陆续推出了多款 AI 芯片，其中 2021 年推出的思元 370 采用了 chiplet 的新技术，整体集成了 390 亿个晶体管，最大算力达到 256TOPS（INT8），也是商用客户里出货量最大、推广最成功的一款产品。公司即将推出的新产品思远 590，性能可对标英伟达 A100，在美国《芯片法案》禁令影响下，该款芯片有望成为国内市场中替代 A100 的主力产品。

表 7：寒武纪比主要产品目录

产品线	产品类型	寒武纪主要产品	推出时间
云端产品线	云端智能芯片及加速卡	思元 100（MLU100）芯片及云端智能加速卡	2018 年
		思元 270（MLU270）芯片及云端智能加速卡	2019 年
		思元 290（MLU290）芯片及云端智能加速卡	2020 年
		思元 370（MLU370）芯片及云端智能加速卡	2021 年
	训练整机	玄思 1000 智能加速器	2020 年
边缘产品线	边缘智能芯片及加速卡	思元 220（MLU220）芯片及边缘智能加速卡	2019 年
IP 授权及软件	终端智能处理器 IP	寒武纪 IA 处理器	2016 年
		寒武纪 IH 处理器	2017 年
		寒武纪 1M 处理器	2018 年
	基础系统软件平台	寒武纪基础软件开发平台（适用于公司所有芯片与处理器产品）	持续研发和升级，以适配新的芯片

资料来源：寒武纪官网，中国银河证券研究院

海光信息：基于 GPGPU 架构的 DCU 产品商业落地。海光信息成立于 2014 年，并于 2019 年切入到 DCU 产品领域，其 DCU 系列产品以 GPGPU 架构为基础，兼容通用的“类 CUDA”环境以及国际主流商业计算软件和人工智能软件，软硬件生态丰富，可广泛应用于大数据处理、人工智能、商业计算等应用领域。DCU 系列产品中的深算一号性能指标堪比国际上同类型高端产品，并在 2021 年实现商业化应用，深海二号正在研发中，也将成为算力芯片市场强有力的竞争者之一。

图 57：海光 DCU 产品深算一号和其他产品的对比

项目	海光	NVIDIA	AMD
品牌	深算一号	Ampere100	MI100
生产工艺	7nmFinFET	8nmFinFET	9nmFinFET
核心数量	4096（64CUs）	2560 CUDA processors 640Tensor processors	120CUs
内核频率	Up to 1.5Ghz（FP64） Up to 1.7Ghz（FP32）	Up to 1.53Ghz	Up to 1.5Ghz（FP64） Up to 1.7Ghz（FP32）
显存容量	32GB HBM2	80GB HBM2e	32GB HBM2
显存位宽	4096 bit	5120bit	4096bit
显存频率	2.0 GHZ	3.2 GHZ	2.4 GHZ
显存带宽	1024 GB/s	2039 GB/s	1228 GB/s
TDP	350W	400W	300W
CPU to GPU 互联	PCIe Gen4 x 16	PCIe Gen4 x 17	PCIe Gen4 x 18
GPU toGPU 互联	xGMI x 2 Up to 184 GB/s	NVLink Up to 600 GB/s	Infinity Fabric x 3 Up to 600 GB/s

资料来源：海光信息招股说明书，与非网，中国银河证券研究院

图 58：海光 DCU 产品形态

资料来源：海光信息招股说明书，中国银河证券研究院

龙芯中科：GPGPU 预计 23 年流片。龙芯中科成立于 2010 年，主营业务为处理器及配套芯片的研制、销售及服务，主要产品与服务包括处理器及配套芯片产品与基础软硬件解决方案业务。上市之初，公司就有 GPGPU 设计技术的储备，并募集资金 10.5 亿投向高性能通用图形处理器芯片及系统研发项目，主要针对图形加速、科学计算尤其是人工智能应用的需求。2022 年 9 月 5 日，龙芯中科在业绩说明会上表示，公司 GPGPU 研发项目进展顺利，将于 2023 年流片，公司有望成为 AI 算力芯片领域新星。

图 59：龙芯中科募投资金使用明细

项目名称	项目投资总额	拟使用募集资金额(单位：万元)
先进制程芯片研发及产业化项目	12576045.00	125760.45
高性能通用图形处理器芯片及系统研发项目	10542645.00	10542645.00
补充流动资金	120000.00	120000.00
合计	35118690.00	35118690.00

资料来源：龙芯中科年报，中国银河证券研究院

图 60：龙芯中科拥有 GPGPU 技术储备

核心技术储备名称	概况	类型
同时多线程技术	单个处理器核支持同时执行两个及两个以上硬线程的技术，支持根据线程业务负载和硬件资源使用情况在单线程和多线程模式间的自动切换，实现单线程绝对性能和多线程任务吞吐率的平衡	关键核心技术研发
新一代系统级虚拟化技术	新一代的系统机虚拟化技术在实现高效安全的 CPU 和内存虚拟化的基础上，进一步优化中断虚拟和 IO 虚拟效率中断和 DMA 可以直接注入虚拟机，无需陷入到宿主环境处理，可以显著提升虚拟化场景下 IO 和多核通信性能；在桥片中，拓展总线协议和设备功能，实现 IO 设备的直通、隔离，以及多队列设备的高效安全处理	关键核心技术研发
GPGPU 设计技术	面向 GPU 超大规模并行处理的特点，设计完整的软硬件框架，适配 GPU 通用计算的需求，优化流处理器结构，不断提高单位面积 / 功耗下的算力密度，提高整机竞争力	关键核心技术研发

资料来源：龙芯中科官网，中国银河证券研究院

（二）先进封装：“后摩尔时代”先进封装突破极限

通富微电：持续突破先进封装技术。通富微电深耕于集成电路封装测试一体化服务，产品覆盖面广且技术全面。近年来，公司积极布局 Chiplet、2.5D/3D、扇出型、圆片级、倒装焊等封装技术，可为客户提供多样化的 Chiplet 封装解决方案，并且已为 AMD 大规模量产 Chiplet 产品。在高性能计算机领域，公司已建成国内顶级 2.5D/3D 封装平台（VISionS）及超大尺寸 FCBGA 研发平台，并且完成高层数再布线技术开发，同时可以为客户提供晶圆级和基板级 Chiplet 封测解决方案。2022 年上半年，公司在 2.5D/3D 先进封装平台方面，再度取得突破性进展，BVR 技术实现通线并完成客户首批产品验证，2 层芯片堆叠的 CoW 技术完成技术验证。依托于丰富的国际市场开发经验和坚实的技术基础，公司有望抓住先进封测市场机遇，稳

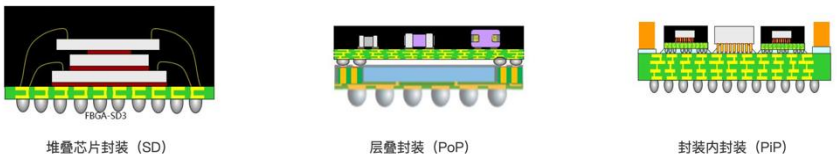
固其行业龙头的地位。

图 61：使用 FCBGA 技术生产的产品

资料来源：通富微电官网，中国银河证券研究院

长电科技：半导体封装行业龙头。长电科技是全球领先的集成电路制造和技术服务提供商，可以提供全方位的芯片成品制造一站式服务，拥有行业领先的半导体先进封装技术（如 SiP、WL-CSP、FC、eWLB、PiP、PoP 及 XDFOITM 系列等）。2021 年公司推出的面向 3D 封装的 XDFOITM 系列产品，为高性能计算领域提供了业界领先的超高密度异构集成解决方案。子公司星科金朋与客户共同开发了基于高密度 Fan out 封装技术的 2.5D fcBGA 产品，同时认证通过 TSV 异质键合 3D SoC 的 fcBGA，提升了集成芯片的数量和性能，为进一步全面开发 Chiplet 所需高密度高性能封装技术奠定了坚实的基础。2022 年，公司推动实施技术开发 5 年规划，包括对 2.5D/3D chiplet，高密度多叠加存储技术等八大类逾三十项先进技术开展前瞻性研发，将进一步推动技术和产品价值进一步提升，持续增强市场竞争力。

图 62：长电科技 2.5D/3D 集成技术解决方案



资料来源：长电科技官网，中国银河证券研究院

（三）服务器 PCB：AI 服务器催动 PCB 技术升级

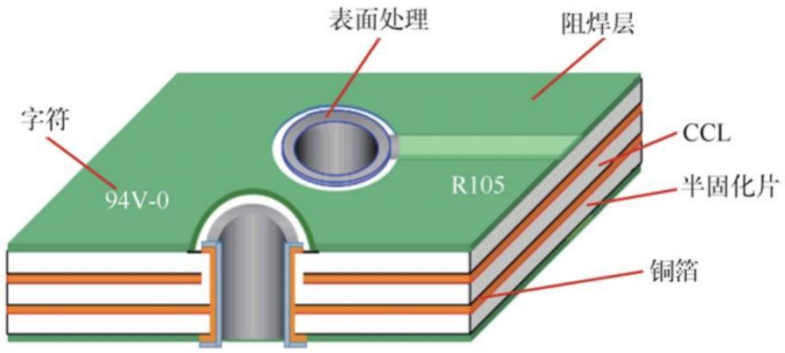
服务器面向数据处理需求迭代，大算力时代引爆 AI 服务器需求。服务器是算力的载体，普通的服务器主要为智能手机、PC 等提供基础的算力和数据存储支持，多以 CPU 为算力的提供者、采用串行架构，无法满足大算力时代不断攀升的数据量引发的数据处理需求。AI 服务器多采用 CPU+GPU/TPU/其他加速卡的异构形式，一般配置四块以上 GPU 卡，可以满足高吞吐量互联的需求，提供强大的算力支持。由 ChatGPT 引爆的 AIGC 场景增多驱动智能算力的规模不断增长，因此人工智能服务器的需求量也将不断攀升。

图 63：华为 Atlas 800 训练服务器内部结构图

资料来源：CNDS，中国银河证券研究院

PCB 是服务器的重要组成部分，技术升级势在必行。服务器算力的提升除依靠 CPU、加速芯片组外，PCLe 总线标准的提升也是必不可少的环节。根据 Intel 规划，服务器平台方案正由 Purely 转为 Whitley，而 Whitley 中的 Ice Lake 方案也将首次支持 PCLe4.0 总线设计，下一代 Eagle Stream 平台将同步支持 PCLe5.0。PCB 是 PCIe 总线中的关键组件，高等级的总线标准需要 PCB 层数和基材的支持，其中 PCB 层数需求将从 3.0 的 8-12 层提升至 5.0 的 16 层以上；CCL 材料的 Df 值也需要同步降低。AI 服务器需求量的提升和 PCB 技术的升级必将带来 PCB 产品的量价齐升。

图 64： PCB 四层板结构示意图



资料来源：深圳无双信息技术有限公司，中国银河证券研究院

沪电股份：高端 PCB 行业龙头。沪电股份深耕 PCB 行业 20 年，在技术、质量、成本、品牌、规模等方面形成相对竞争优势，居行业领先地位。公司坚持差异化竞争战略，重点生产技术含量高、应用领域相对高端的差异化产品。在高性能计算领域，应用于 AI 加速、Graphics 的产品，应用于 GPU、OAM、FPGA 等加速模块类的产品以及应用于 UBB、BaseBoard 的产品已批量出货，目前正在预研应用于 UBB2.0、OAM2.0 的产品。公司持续加大在高端产品领域的研发投入，正在进行的高速 HDI 长期可靠性研究也将强化公司在 AI 加速核心产品市场的竞争力。

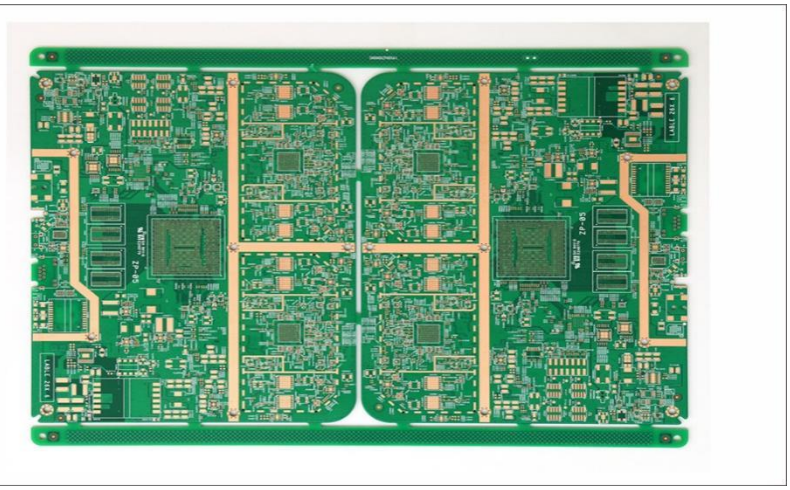
表 8：沪电股份研发项目进度

主要研发项目名称	项目目的	项目进展	拟达到的目标	预计对公司未来发展的影响
低粗糙度高速氧化工艺技术研发	提升高速信号完整性技术能力	已完成	匹配 112Gpbs 电性能技术需求，满足产品可靠性需求，实现技术产品化	提升高速产品的技术优势
高纵横比深微孔技术能力开发	提升高密、Power 通流技术能力		匹配路由器、交换机对高密度、大电流高通流的 PCB 技术要求，实现量产	提升路由、交换产品的技术优势
企业网大尺寸交换网板产品开发	提升企业网市场产品技术能力		实现 48"超大尺寸网板加工技术量产	提升该领域技术领先优势
EGS 等级服务器产品开发	提升数据中心服务器市场竞争力		实现使用 IntelAMD 等新一代服务器产品规模化量产	提高公司在服务器产品市场的竞争力
高速 Low loss 国产替代材料开发	材料选自主可控，提升市场竞争力	进行中	对应不同等级材料均有国产替代材料	提升材料自主选择权
高速 HDI 长期可靠性研究	重算力加速模块及 102.4T Switch/Router 产品可靠性技术预研		深度参与行业客户对产品技术的预研，储备关键核心技术	强化公司企业通讯市场及 AI 加速核心产品市场的竞争优势，提高客户黏着度。

资料来源：沪电股份年报，中国银河证券研究院

胜宏科技：服务器领域应用实现从 0 到 1。胜宏科技成立于 2006 年，主要从事高密度印制线路板的研发、生产和销售，主要产品包括双面板、多层板（HDI）等。2021 年，在消费电子市场疲软的环境下，公司及时调整客户结构和产品结构，并顺利导入通讯、服务器、芯片等多家国内外优质客户。公司坚持优质客户与高端产品的战略布局，建立起了高速 SI 能力系统，支持通讯、服务器高端客户的开发，也开展了“平台服务器主板研发”、“服务器硬盘用高频主板研发”等研发项目，为企业的持续增长注入了活力。

图 65： 胜宏科技服务器用 PCB 板结构图

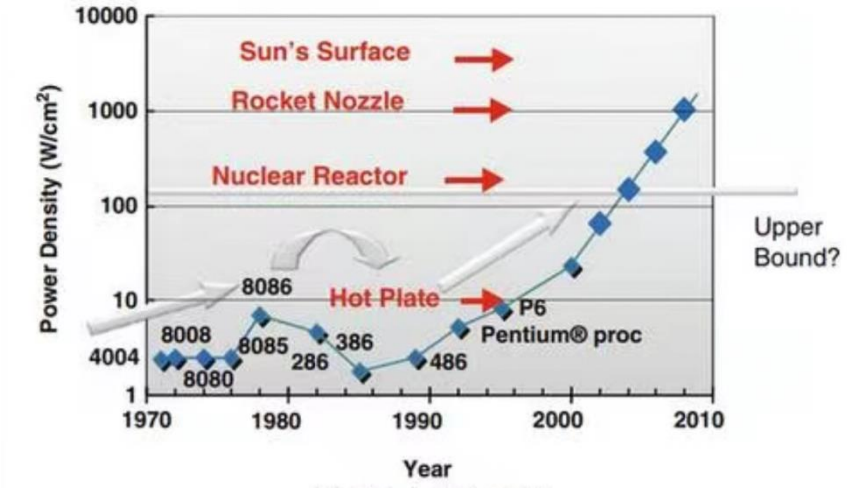


资料来源：胜宏科技官方，中国银河证券研究院

（四）散热：功耗与算力同步提升，散热技术面临挑战

芯片工作温度显著影响性能。芯片算力不断提升的背后是计算效率的提升和功耗的增加。芯片功耗的增加会使得芯片温度升高，而分子热运动也会随着温度升高而增大，影响到载流子的定向迁移，使芯片的漏电流及电流增益加大，从而增大芯片的功耗，形成恶性循环。AI 服务器尤为注重纯算力的运算，因此温度升高，AI 服务器降频运行现象尤为明显，散热技术的升级势在必行。

图 66：芯片的功率密度近年来不断提升

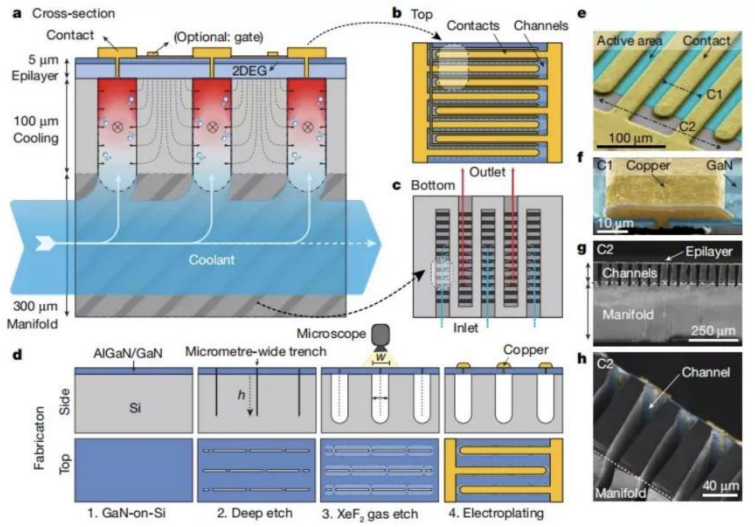


资料来源：Intel，中国银河证券研究院

散热技术向液冷和芯片级演进。在功耗提升的同时，芯片整体尺寸也越来越小，电子芯片工作过程中所呈现出的热流密度大幅提升，传统的风冷散热已经很难满足当下电子芯片的散热需求，因此，以导热性能是空气 15-25 倍的液体作为冷却介质将成为未来散热技术的主要发展方向之一。除冷却介质外，散热部分和核心发热源距离的不同也会影响散热效果。随着散热技术的升级，目前散热方案正在从房间级、机柜级、服务器级向着芯片级演进。在芯片级液冷

技术、相变储热散热技术、蒸发冷却技术这三种芯片级散热方案中，芯片级液冷技术散热性能好、散热效率高、能耗小、占地空间小、可靠性强，因此将逐渐成为 AI 服务器主流散热方案。

图 67：协同设计的微流体冷却电子设备结构图及各角度试图



资料来源：《Nature》，中国银河证券研究院

中石科技：热管理解决方案产品可应用于服务器/数据中心。中石科技成立于 1997 年，公司基于为全球龙头通信设备供应商提供热管理解决方案二十余年的经验，不断丰富产品矩阵，拓宽下游应用场景。在服务器/数据中心领域，公司提供的主要产品：热模组（尤其是液冷散热模组）、导热垫片、导热硅脂、导热凝胶、导热相变材料、导热碳纤维垫等；公司目前已向国内外多家上述终端应用企业批量供货。公司宜兴募投项目的水冷和液冷散热模组等产品已逐步落地，有望今年交付，将进一步提升公司在服务器/数据中心应用领域的竞争力。

表 9：中石科技热解决方案产品

产品名称	细分产品	特点及行业地位	应用场景
高导热石墨产品	人工合成石墨、天然石墨、石墨烯高导热膜、单体厚石墨导热膜、多层复合石墨导热膜等。	根据日本富士经济出版的报告， 公司是人工合成高导热石墨膜全龙头公司，品类齐全，技术领先。	手机、平板电脑、充电模组、VR/AR、智能家居设备、汽车电子、新能源逆变器、新型显示装置、高功率电力电了等。

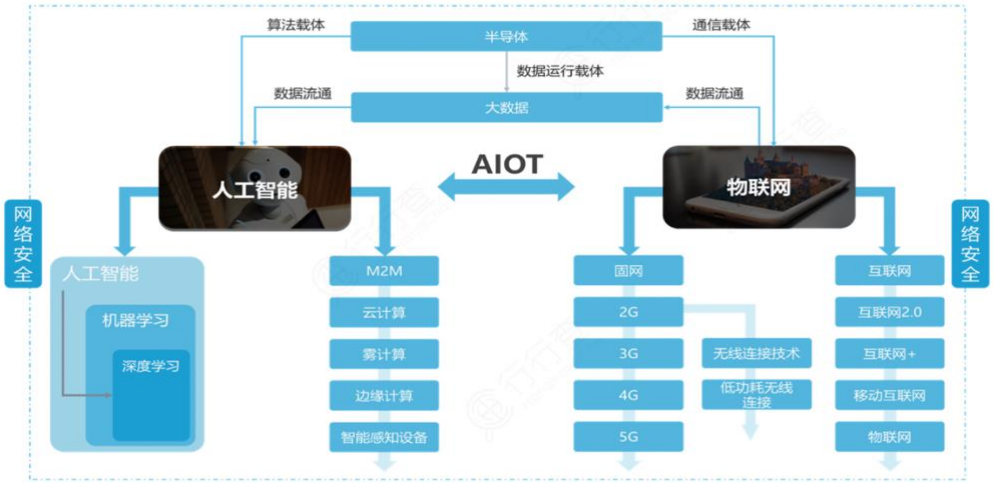
导热界面材料 TIM	导热填隙垫片、导热凝胶、导热硅脂、相变材料、储热材料、高回弹石墨材料、界面石墨产品等。	在导热界面材料领域，公司深耕行业 16 年，是全球通信行业、消费类电子主流导热界面材料供应商，公司多项产品属于业内首创。	通信基站、手机、平板电脑、智能家居设备、汽车电子（三电系统）、电装设备等。
热管	标准热管、薄型热管、超薄热管、大功率薄型热管 HPS 等。	用于热远点传播，特点是高效导热、灵活应用，用于大功率芯片及散热空间小的产品。	笔记本、服务器、游戏机、VR/AR、通信设备等。
均热板	标准均热板、薄型均热板、超薄均热板等。	用于热面传播，特点是超薄（最薄可达 0.25mm）、低热阻、高散热、多向散热。	手机、平板、新能源等。
热模组	风冷散热模组-服务器散热模组、笔电散热模组、清洁能源散热模组、其他定制化散热模组等；液冷散热模组-管式液冷板、埋管式液冷板、一体式液冷板等。	风冷散热模组-散热功率高；液冷散热模组-防水防尘设计、 比风冷更节能、热流密度好、可靠好、可以进行灵活的流体通道设计，适应更高散热功率场景。	服务器/数据中心、笔记本、PC、一体式电脑、游戏机、投影仪、医疗、电子、电力等。

资料来源：中石科技，中国银河证券研究院

（五）AIoT：从“万物互联”到“万物智联”

AI 技术可以赋予 LoT“人工智能大脑”。人工 ALoT 即“AI+IoT”，指的是人工智能技术与物联网在实际应用中的落地融合。物联网可以将人与物、物与物连接成为一个整体，通过 LoT 智能设备生成海量数据；AI 技术可以对海量数据进行深度学习、判断用户的习惯，提升用户体验，两者相辅相成，推动“万物互联”向“万物智联”进化。ChatGPT 的出现使得人工智能技术在语言交互方面的应用更为广泛，近日推出的插件功能，将进一步促进 AI 技术和其他产业的融合，AIoT 产业也将在 AI 技术升级的推动下不断发展。

图 68：AIoT 技术架构图



资料来源：亿欧智库，中国银河证券研究院

瑞芯微：中国领先的 AIoT 芯片设计公司。瑞芯微成立于 2001 年，专注于集成电路设计和研发。近几年，公司跟随市场趋势变化，大力研发 AIoT 产品、开拓相关市场，积极打造 AIoT 生态，已经成为国内领先的 AIoT 芯片供应商。公司 AIoT 旗舰芯片 RK3588 系列是目前国内

顶配高端 AIoT 芯片，可以应用于 ARM PC、平板、高端摄像头、NVR、8K 和大屏设备、汽车智能座舱、云服务设备及边缘计算、AR/VR 等八大方向市场。RK3588 的成功量产，也意味着瑞芯微 AIoT 大厦的基本成型，AIoT 业务将成为未来营收增长的主力军。

表 10：瑞芯微部分 AIoT 产品及其应用领域

类别	子类	主要特点	主要产品系列	主要应用领域
智能应用处理器芯片	高性能应用处理器	采用高性能 CPU 和 GPU 内核，新一代芯片还增加了 NPU，具有强大的多媒体处理能力，以及众多外设接口，可以适应众多复杂场景应用的需求，可以运行 Android、Linux 等操作系统，是公司的代表性旗舰产品	RK3588 系列	ARM PC、平板、高端摄像头、NVR、8K 和大屏设备、汽车智能座舱、云服务设备及边缘计算、AR/VR 等
			RK3399 系列	刷脸识别及支付、ARM 服务器、视频会议系统、商业显示、行业平板和电子白板、自助设备、开发板及工控等
			RK3288 系列	商业显示、收银机、刷脸识别及测温、行业平板、开发板及工控、自助设备、云终端、电纸书、汽车电子、视频会议系统等
			RK3568/RK3566 系列	平板电脑、NVR、NAS、电纸书、云终端、网关等
	通用应用处理器	具有适当的处理能力，性价比高	RK3368 系列	教育电子、收银机、智能家电、智能门禁等

资料来源：瑞芯微，中国银河证券研究院

全志科技：中国领先的 AIoT 芯片设计公司。全志科技成立于 2007 年，是卓越的智能应用处理器 SoC、高性能模拟器件和无线互联芯片设计厂商。在 AIoT 领域，公司与行业头部一线智能音箱标杆客户保持产业深度合作，R 系列芯片产品已实现带屏、无屏音箱全面量产。基于智能语音的技术积累及生态布局，公司也与智能家电、扫地机器人、陪伴机器人、AI 教育（学习机、词典笔）等领域重要客户深度合作，推出了 MR 系列、V853 芯片等多款产品，丰富了在 AIoT 领域的产品矩阵。

图 69：全志科技 R 系列产品图谱

资料来源：全志科技官网，中国银河证券研究院

四、投资建议

我们认为，在 ChatGPT 等应用商业化出现落地方式，AIGC 创作内容不断增长的条件下，芯片作为 AI 行业的基础设计，为 AI 训练和数据计算提供支持，未来 AI 应用落地层面对庞大算力的需求更为重要，因此，相关算力产业链未来发展值得期待。我们看好国内相关算力产业链公司的未来发展，建议关注：

GPU、加速卡、AI 芯片：寒武纪(688256.SH)、景嘉微(300474.SZ)、海光信息(688041.SH)

先进封装：通富微电（002156.SZ)、长电科技（600584.SH)、深科技（000021.SZ)

服务器及加速卡 PCB：沪电股份（002463.SZ)、胜宏科技（300476.SZ)

AIoT 产业链：瑞芯微（603893.SH)、全志科技（300458.SZ)、晶晨股份（688099.SH)、富瀚微（300613.SZ)

芯片 IP：芯原股份（688521.SH)、华大九天（301269.SZ)

存储芯片/模组/PCIe：兆易创新(603986.SH)、江波龙(301308.SZ)、北京君正(300223.SZ)、聚辰股份（688123.SH)、澜起科技（688008.SH)

散热材料：中石科技（300684.SZ)、飞荣达（300602.SZ)

建议关注

股票名称	股票代码	当前价格	EPS(元)			PE（X）		
			2011E	2012E	2013E	2011E	2012E	2013E
寒武纪-U	688256.SH	183.00	-1.89	-1.09	-	-	-	-
景嘉微	300474.SZ	113.05	0.63	0.93	1.27	178.12	121.55	89.10
海光信息	688041.SH	66.25	0.61	0.91	-	107.88	72.49	-
通富微电	002156.SZ	24.79	0.39	0.74	1.06	63.08	33.40	23.45
长电科技	600584.SH	33.40	1.85	2.04	2.37	18.10	16.40	14.07
深科技	000021.SZ	15.34	0.53	0.66	0.74	28.81	23.17	20.81
沪电股份	002463.SZ	21.60	0.88	1.12	1.42	24.50	19.34	15.21
胜宏科技	300476.SZ	19.50	0.99	1.22	1.54	19.76	15.94	12.62
瑞芯微	603893.SH	95.85	0.93	1.46	2.07	103.48	65.67	46.41
全志科技	300458.SZ	28.90	0.54	0.69	-	53.47	42.10	-
晶晨股份	688099.SH	88.38	2.78	3.81	-	31.77	23.20	-
富瀚微	300613.SZ	74.31	1.85	2.40	3.12	40.24	30.90	23.81
华大九天	301269.SZ	122.34	0.34	0.47	0.64	355.33	259.25	190.92
兆易创新	603986.SH	117.90	3.72	4.03	4.98	31.68	29.22	23.67
江波龙	301308.SZ	83.77	0.96	1.38	1.89	87.31	60.70	44.44
北京君正	300223.SZ	85.43	1.94	2.44	3.16	44.15	35.05	27.07
聚辰股份	688123.SH	103.93	4.70	6.17	-	22.12	16.84	-
澜起科技	688008.SH	70.17	1.64	2.37	-	42.75	29.65	-
中石科技	300684.SZ	18.97	0.69	0.97	1.35	27.52	19.56	14.07

飞荣达	300602.SZ	17.35	0.07	0.40	0.80	250.00	43.09	21.75

资料来源：Wind, 中国银河证券研究院

五、风险提示

1.行业需求不及预期的风险：市场需求的变化和市场份额的变化可能会影响 AI 行业的需求变化，若相关企业对行业的需求波动反应不敏感，可能会造成企业盈利不及预期的情况。

2.应用落地不及预期的风险：AI 行业更关注行业最后英应用落地的场景，相关企业的山歌优化进程推动不及预期，可能会造成公司的产品或技术不符合市场需求，导致行业内公司收入无法创造回报的风险。

3.政策风险：目前国内 AI 企业与海外 GPU 芯片合作存在障碍，未来政策趋势尚未确定，可能会对 AI 企业相关基础设计的供应存在不稳定性和价格波动的影响，对公司未来研发进度和产品进度造成影响，进而对公司造成不确定性的风险。

4.下游技术迭代不及预期：存在行业技术迭代速度不及预期从而对需求造成影响的风险。

插 图 目 录

表 格 目 录

分析师简介及承诺

高峰，北京邮电大学电子与通信工程硕士，吉林大学工学学士。2 年电子实业工作经验，6 年证券从业经验，曾就职于渤海证券、国信证券、北京信托证券部。2022 年加入中国银河证券研究院，担任电子团队组长，主要从事硬科技方向研究。

王子路，英国布里斯托大学金融与投资硕士，山东大学经济学学士，2 年科技产业研究经验，2020 年加入中国银河证券研究院，从事电子行业研究。

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。

评级标准

行业评级体系

未来 6-12 个月，行业指数（或分析师团队所覆盖公司组成的行业指数）相对于基准指数（交易所指数或市场中主要的指数）

推荐：行业指数超越基准指数平均回报 20%及以上。

谨慎推荐：行业指数超越基准指数平均回报。

中性：行业指数与基准指数平均回报相当。

回避：行业指数低于基准指数平均回报 10%及以上。

公司评级体系

推荐：指未来 6-12 个月，公司股价超越分析师（或分析师团队）所覆盖股票平均回报 20%及以上。

谨慎推荐：指未来 6-12 个月，公司股价超越分析师（或分析师团队）所覆盖股票平均回报 10%－20%。

中性：指未来 6-12 个月，公司股价与分析师（或分析师团队）所覆盖股票平均回报相当。

回避：指未来 6-12 个月，公司股价低于分析师（或分析师团队）所覆盖股票平均回报 10%及以上。

免责声明

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：苏一耘 0755-83479312 suyiyun_yj@chinastock.com.cn

程 曦 0755-83471683 chengxi_yj@chinastock.com.cn

上海地区：何婷婷 021-20252612 hetingting@chinastock.com.cn

陆韵如 021-60387901 luyunru_yj@chinastock.com.cn

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