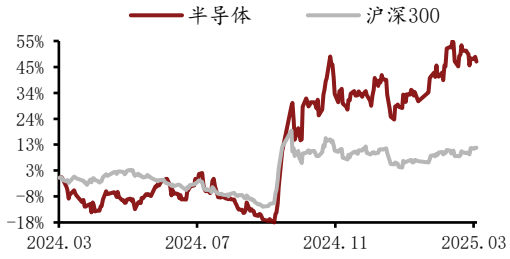
AI 算力芯片是“AI 时代的引擎”，河南省着力布局

分析师：邹臣
登记编码：S0730523100001
zouchen@ccnew.com 021-50581991

证券研究报告-行业深度分析

发布日期：2025 年 03 月 20 日

半导体相对沪深 300 指数表现



资料来源：聚源，中原证券

相关报告

《半导体行业月报：国内 RISC-V 生态加速发展，存储器价格有望逐步回升》 2025-03-10

《半导体行业月报：美国半导体出口管制进一步升级，DeepSeek 热潮有望推动端侧 AI 发展》 2025-02-10

《半导体行业月报：豆包 AI 生态加速发展，关注国内 AI 算力产业链》 2025-01-10

联系人：李智
电话： 0371-65585629
地址： 郑州郑东新区商务外环路 10 号 18 楼
地址： 上海浦东新区世纪大道 1788 号 T1 座 22 楼

报告要点：

- **AI 算力芯片是“AI 时代的引擎”。**ChatGPT 热潮引发全球科技企业加速布局 AI 大模型，谷歌、Meta、百度、阿里巴巴、华为、DeepSeek 等随后相继推出大模型产品，并持续迭代升级；北美四大云厂商受益于 AI 对核心业务的推动，持续加大资本开支，国内三大互联网厂商不断提升资本开支，国内智算中心加速建设，推动算力需求高速成长。人工智能进入算力新时代，全球算力规模高速增长，根据 IDC 的预测，预计全球算力规模将从 2023 年的 1397 EFLOPS 增长至 2030 年的 16 ZFLOPS，预计 2023-2030 年复合增速达 50%。AI 服务器是支持生成式 AI 应用的核心基础设施，AI 算力芯片为 AI 服务器提供算力的底层支撑，是算力的基石。AI 算力芯片作为“AI 时代的引擎”，有望畅享 AI 算力需求爆发浪潮，并推动 AI 技术的快速发展和广泛应用。

- **AI 算力芯片以 GPU 为主流，定制 ASIC 芯片市场高速成长。**AI 算力芯片按应用场景可分为云端、边缘端、终端 AI 算力芯片，本文主要针对于云端 AI 算力芯片。根据芯片的设计方法及应用，AI 算力芯片可分为通用型 AI 芯片和专用型 AI 芯片，当前 AI 算力芯片以 GPU 为主流。随着 AI 算力规模的快速增长将催生更大的 GPU 芯片需求，根据 Statista 的数据，2023 年全球 GPU 市场规模为 436 亿美元，预计 2029 年市场规模将达到 2742 亿美元，预计 2024-2029 年复合增速达 33.2%。根据 TechInsights 的数据，2023 年英伟达在数据中心 GPU 出货量中占据 98%的市场份额，主导全球 GPU 市场。GPU 生态体系复杂，建设周期长、难度大，GPU 生态体系建立极高的行业壁垒。AI ASIC 是一种专为人工智能应用设计的定制集成电路，具有高性能、低功耗、定制化、低成本等特点。由于英伟达垄断全球数据中心 GPU 市场，云厂商为了提升议价能力及供应链多元化，推动数据中心定制 ASIC 芯片市场高速成长，预计增速快于通用 AI 算力芯片。根据 Marvell 的数据，2023 年数据中心定制 ASIC 芯片市场规模约为 66 亿美元，预计 2028 年市场规模将达到 429 亿美元，预计 2023-2028 年复合增速达 45%。近年来美国不断加大对高端 GPU 的出口管制，国产 AI 算力芯片厂商迎来黄金发展期。

- **DeepSeek 有望推动国产 AI 算力芯片加速发展。**DeepSeek 通过技术创新实现大模型训练及推理极高性价比，DeepSeek 模型的技术创新主要体现在采用混合专家（MoE）架构、多头潜在注意力机制（MLA）、FP8 混合精度训练技术、多 Token 预测（MTP）及蒸馏技术等。DeepSeek-V3 性能对标 GPT-4o，DeepSeek-R1 性能对标 OpenAI o1；根据 DeepSeek 在 2025 年 1 月 20 日公布的数据，DeepSeek-R1 API 调用成本不到 OpenAI o1 的 5%。DeepSeek-R1 实现模型推理极高性价比，蒸馏技术使小模型也具有强大的推理能力及低成本，将助力 AI 应用大规模落地，并有望推动推理需求加速释放。IDC 预计 2028 年中国 AI 服务器用于推理工作负载占比将达到 73%，由于推理服务器占比远高于训练服

务器，用于推理的 AI 算力芯片国产替代空间更为广阔。国产算力生态链已全面适配 DeepSeek，DeepSeek 通过技术创新提升 AI 算力芯片的效率，进而加快国产 AI 算力芯片自主可控的进程，国产 AI 算力芯片厂商有望加速发展，并持续提升市场份额。

- **河南省着力布局 AI 算力芯片，产业链初具雏形。**河南省将算力作为支撑数字河南建设的重要底座和驱动数字化转型的新引擎，致力于打造面向中部、辐射全国的算力调度核心枢纽和全国重要的算力高地。河南省的算力产业布局以“一核四极多点”为核心框架，以郑州市（含航空港区）为核心，支持洛阳、鹤壁、商丘、信阳等城市作为区域增长极。河南省依托省内先进计算企业，积极引进芯片等上游企业，吸引集聚服务器操作系统、数据库、中间件开发骨干企业，打造先进计算产业园，构建算力产业生态。龙芯中科在鹤壁建设的芯片封装基地已正式投产，并在郑州设立中原总部基地，中原总部基地将建设研发创新中心、生态适配中心、信创展示中心等；河南投资集团通过基金投资沐曦集成，推动沐曦集成在河南落地；河南省政策大力扶持 AI 算力芯片产业，通过引进、投资、培育本土企业等方式布局 AI 算力芯片，产业链初具雏形。

- **相关企业。**河南省 AI 算力芯片产业相关企业主要有龙芯中科、沐曦等。

风险提示：国际地缘政治冲突加剧风险，下游需求不及预期风险，市场竞争加剧风险，新产品研发进展不及预期风险，国产替代进展不及预期风险。

内容目录

1. AI 算力芯片是“AI 时代的引擎”..... 5
1.1. 大模型持续迭代，推动全球算力需求高速成长 5
1.2. AI 算力芯片是算力的基石 8
2. AI 算力芯片以 GPU 为主流，定制 ASIC 芯片市场高速成长..... 9
2.1. AI 算力芯片可应用于云端、边缘端、终端，当前以 GPU 为主流 9
2.2. 英伟达主导全球 GPU 市场，GPU 生态体系建立极高的行业壁垒 12
2.3. 云厂商推动定制 ASIC 芯片市场高速成长 18
2.4. 美国不断加大对高端 AI 算力芯片出口管制，国产厂商迎来黄金发展期 22
3. DeepSeek 有望推动国产 AI 算力芯片加速发展..... 24
4. 河南省着力布局 AI 算力芯片，产业链初具雏形..... 30
5. 河南省 AI 算力芯片产业相关企业..... 33
5.1. 龙芯中科 33
5.2. 沐曦 34

图表目录

图 1：全球部分科技企业发布大模型产品情况 5
图 2：GPT-4.5 与人类测试者的对比评估情况 5
图 3：GPT-4o SimpleQA 性能对比情况 5
图 4：o3 在 SWE-benchVerified、Codeforces 测试中表现优于 o1 6
图 5：o3 在 GPQA 测试中大幅优于 o1 6
图 6：2012-2023 年各领域重要的机器学习模型训练算力需求情况 6
图 7：2020-2024 年北美四大云厂商资本开支情况（亿美元） 7
图 8：2021-2024 年国内三大互联网厂商资本开支情况（百万元） 7
图 9：2019-2030 年全球算力规模情况及预测（EFLOPS） 8
图 10：2019-2026 年中国智能算力市场规模预测 8
图 11：人工智能系统产业链结构图 8
图 12：AI 服务器内部结构图 8
图 13：2023-2028 年全球生成式人工智能和非生成式人工智能服务器市场规模及预测 9
图 14：2024-2028 年中国 AI 服务器市场规模及预测 9
图 15：2018 年服务器成本构成情况 9
图 16：CPU+GPU 异构计算系统方案框图 9
图 17：AI 处理的重心正在从云端向边缘转移 10
图 18：英伟达 A100 GPU 内部架构图 11
图 19：谷歌 TPU 内部架构图 11
图 20：2024 年上半年中国 AI 芯片市场份额情况 11
图 21： AI 算力芯片产业链结构图 12
图 22： GPU 与 CPU 内部架构对比图 12
图 23：GPU 的计算架构 13
图 24：GPU 的内存架构 13
图 25：英伟达多 GPU 系统架构图 14
图 26：英伟达 NVLink 技术演进情况 14
图 27：GPU 应用场景广泛 15
图 28：2023-2029 全球 GPU 市场规模情况及预测（亿美元） 15
图 29：2023 年全球数据中心 GPU 市场竞争格局情况 16
图 30：24Q4 全球 PC GPU 市场竞争格局情况 16
图 31：英伟达数据中心平台 GPU 生态体系架构图 16
图 32：英伟达 CUDA 生态系统的组成 17
图 33：英伟达 CUDA 加速计算解决方案 17
图 34：Marvell 用于数据中心的 ASIC 解决方案 18
图 35：博通 AI ASIC 内部架构图 18

图 36：华为昇腾 AI 生态系统架构图 19
图 37：2023-2028 年数据中心 AI 算力芯片市场规模及预测情况 19
图 38：2023-2028 年数据中心 ASIC 定制芯片市场规模及预测情况 19
图 39：博通累积的定制芯片设计经历 20
图 40：博通定制技术能力与 IP 核情况 20
图 41：TPU 内部架构图 20
图 42：在 TPU v5e 和 v6 Trillium 上运行的 steptime 的 Google 基准测试情况 21
图 43：在 TPU v5e 和 v6 Trillium 上进行 SDXL 基准测试情况 21
图 44：谷歌 TPU 芯片通过 ICI 相互连接 22
图 45：由 TPU v4 建立的算力集群示意图 22
图 46：昇腾计算系统架构框图 24
图 47：昇腾计算产业生态图 24
图 48：DeepSeek-V3 基本架构图 25
图 49：DeepSeek-V3 MTP 应用示意图 25
图 50：DeepSeek-V3 FP8 混合精度框架示意图 26
图 51：DeepSeek-V3 多项评测成绩对标 GPT-4o 26
图 52：DeepSeek-V3 多项评测成绩与其他大模型对比情况 26
图 53：DeepSeek-R1-Zero 的思考时间持续提升以解决推理任务 27
图 54：DeepSeek-R1-Zero、R1、蒸馏小模型的开发流程图 27
图 55：DeepSeek-R1 多项评测成绩对标 OpenAI o1 27
图 56：DeepSeek-R1 蒸馏 32B 和 70B 模型多项评测成绩对标 OpenAI o1-mini 27
图 57：DeepSeek-V3 模型性价比处于最优范围 28
图 58：DeepSeek-R1 与 OpenAI o1 类推理模型 API 定价对比情况（2025 年 1 月 20 日） 28
图 59：2024-2028 年中国 AI 服务器工作负载预测情况 29
图 60：河南省“一核四极多点”算力产业布局示意图 30
图 61：河南空港智算中心示意图 31
图 62：算力将赋能千行百业 31
图 63：超聚变研发中心及总部基地 31
图 64：超聚变稳居中国服务器市场第二 31
图 65：龙芯中科中原总部 32
图 66：联想沐曦 DeepSeek 一体机 32

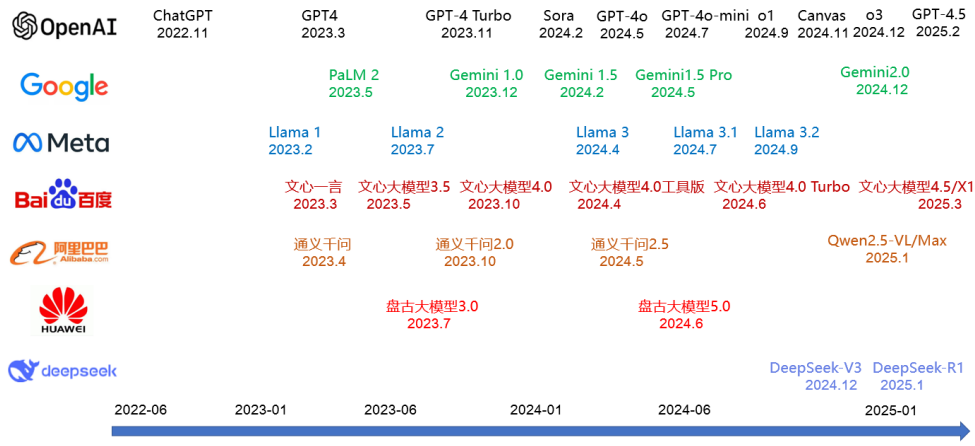
表 1：云端、边缘端、终端应用场景对 AI 算力芯片的算力需求情况 10
表 2：GPU 硬件性能评价参数 13
表 3：英伟达 GeForce 系列 GPU 硬件性能参数对比情况 14
表 4：AI ASIC 与 GPU 性能参数对比情况 18
表 5：谷歌 TPU 历代产品性能参数情况 21
表 6：近年美国对 AI 算力芯片相关制裁政策情况 22
表 7：部分国产 AI 算力芯片技术指标与国际主流产品对比情况 23
表 8：官宣支持 DeepSeek 模型的国产 AI 芯片企业动态 29
表 9：2021-2024 年河南省人工智能产业部分重要产业政策情况 33

1. AI 算力芯片是“AI 时代的引擎”

1.1. 大模型持续迭代，推动全球算力需求高速成长

ChatGPT 热潮引发全球科技企业加速迭代 AI 大模型。ChatGPT 是由美国初创公司 OpenAI 开发、在 2022 年 11 月发布上线的人工智能对话机器人，ChatGPT 标志着自然语言处理和对话 AI 领域的一大步。ChatGPT 上线两个月后月活跃用户数突破 1 亿，是历史上用户增长速度最快的消费级应用程序。ChatGPT 热潮引发全球科技企业加速布局，谷歌、Meta、百度、阿里巴巴、华为、DeepSeek 等科技企业随后相继推出 AI 大模型产品，并持续迭代升级。

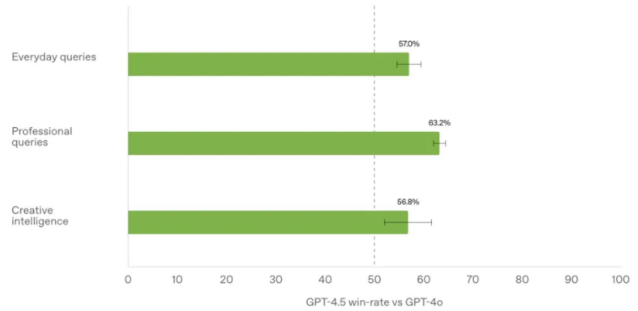
图 1：全球部分科技企业发布大模型产品情况



资料来源：各公司官网，中原证券研究所

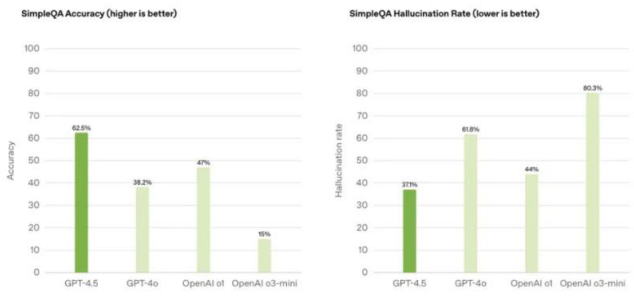
GPT-4.5 带来更自然的交互体验。2025 年 2 月 27 日，OpenAI 正式发布 AI 大模型 GPT-4.5。作为 OpenAI 迄今为止规模最大、知识最丰富的模型，GPT-4.5 在 GPT-4o 的基础上进一步扩展了预训练，与专注于科学、技术、工程和数学(STEM)领域的其他模型不同，GPT-4.5 更全面、更通用。在与人类测试者的对比评估中，GPT-4.5 相较于 GPT-4o 的胜率（人类偏好测试）更高，包括但不限于创造性智能（56.8%）、专业问题（63.2%）以及日常问题（57.0%）；GPT-4.5 带来更自然、更温暖、更符合人类的交流习惯。GPT-4.5 的知识面更广，对用户意图的理解更精准，情绪智能也有所提升，因此特别适用于写作、编程和解决实际问题，同时减少了幻觉现象。

图 2：GPT-4.5 与人类测试者的对比评估情况



资料来源：OpenAI 官网，腾讯，中原证券研究所

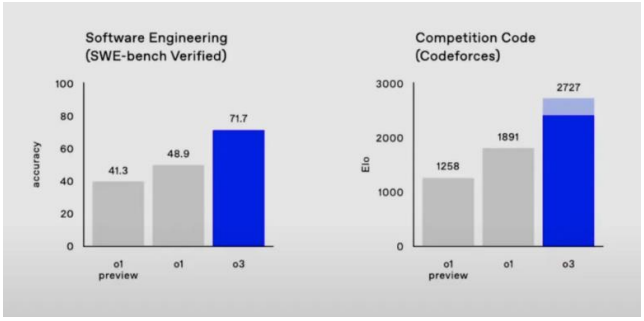
图 3：GPT-4o SimpleQA 性能对比情况



资料来源：OpenAI 官网，腾讯，中原证券研究所

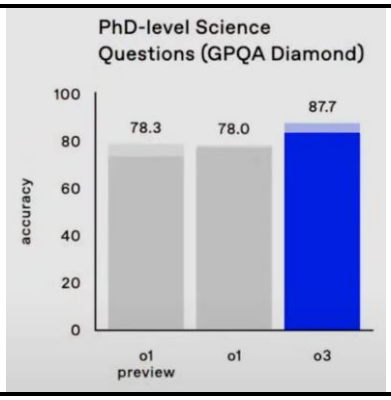
OpenAI o3 进一步提升复杂推理能力。2024 年 12 月 20 日，OpenAI 发布全新推理大模型 o3，o3 模型在多个标准测试中的表现均优于 o1，进一步提升复杂推理能力，在一些条件下接近通用人工智能（AGI）。在软件基准测试（SWE-benchVerified）中，o3 的准确率达到了 71.7%，相较 o1 提升超过 20%；在编程竞赛（Codeforces）中，o3 的评分达到 2727，接近 OpenAI 顶尖程序员水平；而在数学竞赛（AIME）中，o3 的准确率高达 96.7%，远超 o1 的 83.3%；在博士生级别问题测试集（GPQA）中，o3 达到 87.7 分，远超人类选手的程度；在 ARC-AGI 测试中，o3 首次突破了人类水平的门槛，达到 87.5%。

图 4：o3 在 SWE-benchVerified、Codeforces 测试中表现优于 o1



资料来源：OpenAI 官网，腾讯，中原证券研究所

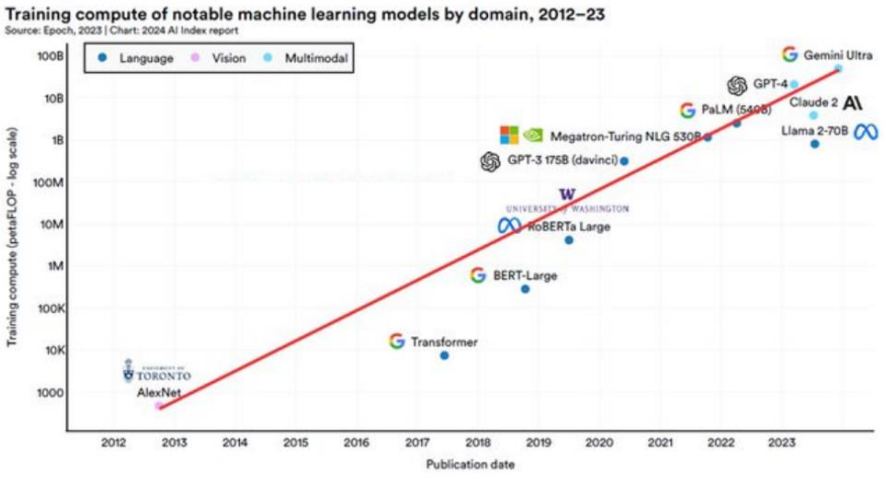
图 5：o3 在 GPQA 测试中大幅优于 o1



资料来源：OpenAI 官网，腾讯，中原证券研究所

大模型持续迭代，推动算力需求高速成长。Scaling law 推动大模型持续迭代，根据 Epoch AI 的数据，2012-2023 年大模型训练的算力需求增长近亿倍，目前仍然在大模型推动算力需求高速成长的趋势中。

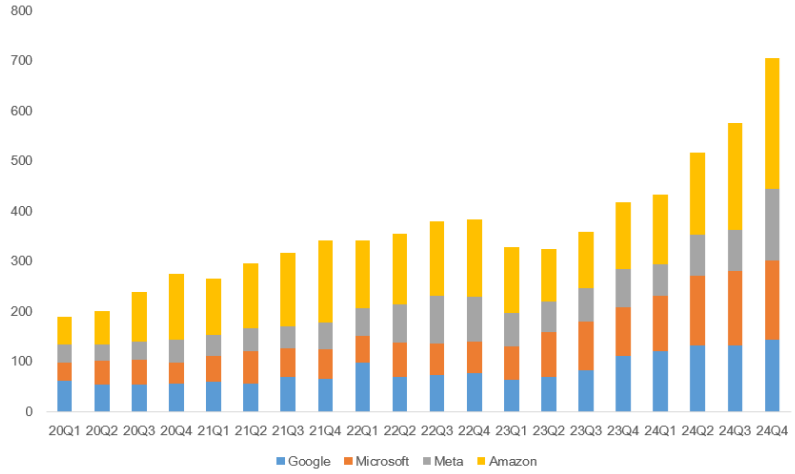
图 6：2012-2023 年各领域重要的机器学习模型训练算力需求情况



资料来源：Epoch AI，网易，中原证券研究所

北美四大云厂商受益于 AI 对核心业务的推动，持续加大资本开支。受益于 AI 对于公司核心业务的推动，北美四大云厂商谷歌、微软、Meta、亚马逊 2023 年开始持续加大资本开支，2024 年四季度四大云厂商的资本开支合计为 706 亿美元，同比增长 69%，环比增长 23%。目前北美四大云厂商的资本开支增长主要用于 AI 基础设施的投资，并从 AI 投资中获得了积极回报，预计 2025 年仍有望继续大幅增加资本开支。

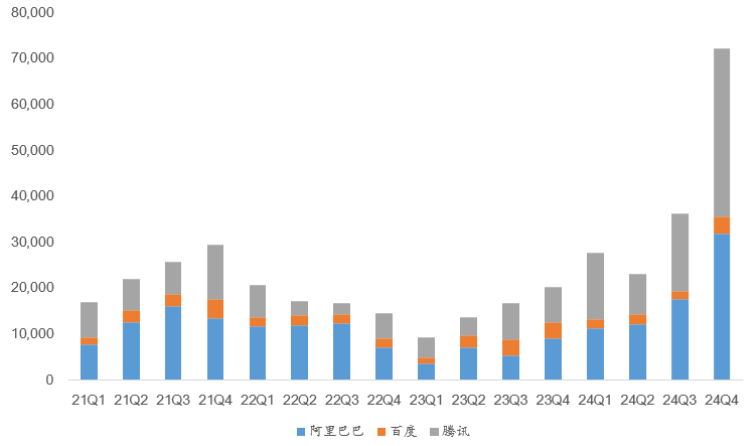
图 7：2020-2024 年北美四大云厂商资本开支情况（亿美元）



资料来源：各公司公告，Wind，中原证券研究所

国内三大互联网厂商不断提升资本开支，国内智算中心加速建设。国内三大互联网厂商阿里巴巴、百度、腾讯 2023 年也开始不断加大资本开支，2024 年四季度三大互联网厂商的资本开支合计为 720 亿元，同比增长 259%，环比增长 99%，预计 2025 年国内三大互联网厂商将继续加大用于 AI 基础设施建设的资本开支。根据中国电信研究院发布的《智算产业发展研究报告(2024)》的数据，截至 2024 年 6 月，中国已建和正在建设的智算中心超 250 个；目前各级政府、运营商、互联网企业等积极建设智算中心，以满足国内日益增长的算力需求。

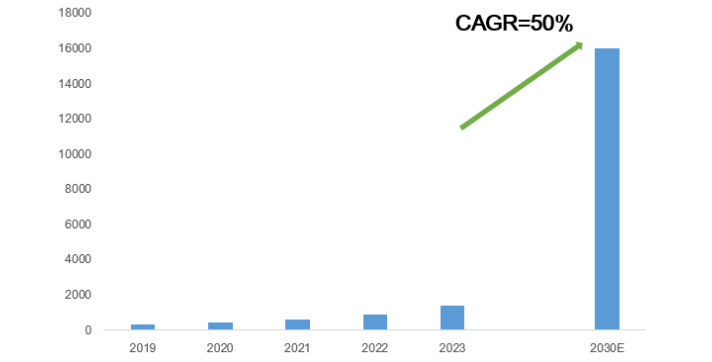
图 8：2021-2024 年国内三大互联网厂商资本开支情况（百万元）



资料来源：各公司公告，中原证券研究所

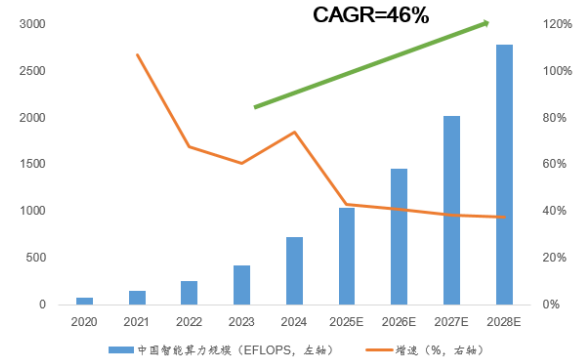
人工智能进入算力新时代，全球算力规模高速增长。随着人工智能的快速发展以及 AI 大模型带来的算力需求爆发，算力已经成为推动数字经济飞速发展的新引擎，人工智能进入算力新时代，全球算力规模呈现高速增长态势。根据 IDC、Gartner、TOP500、中国信通院的预测，预计全球算力规模将从 2023 年的 1397 EFLOPS 增长至 2030 年的 16 ZFLOPS，预计 2023-2030 年全球算力规模复合增速达 50%。根据 IDC 的数据，2024 年中国智能算力规模为 725.3 EFLOPS，预计 2028 年将达到 2781.9 EFLOPS，预计 2023-2028 年中国智能算力规模的复合增速为 46.2%。

图 9：2019-2030 年全球算力规模情况及预测（EFLOPS）



资料来源：IDC，Gartner，TOP500，中国信通院，先进计算暨算力发展指数蓝皮书（2024 年），中原证券研究所

图 10：2019-2026 年中国智能算力市场规模预测

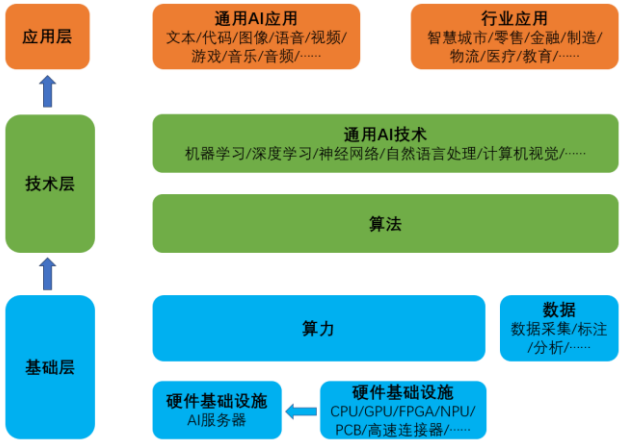


资料来源：IDC，2025 年中国人工智能计算力发展评估报告，中原证券研究所

1.2. AI 算力芯片是算力的基石

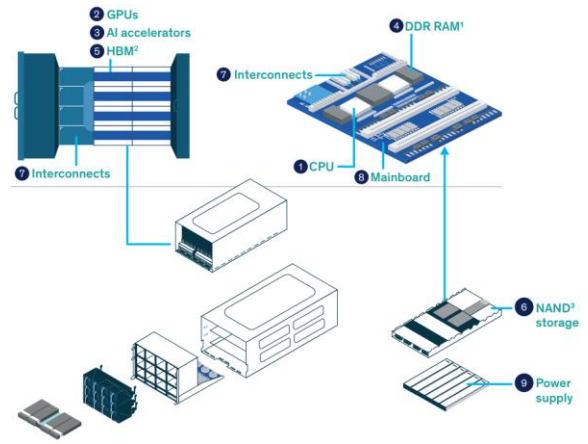
AI 服务器是支撑生成式 AI 应用的核心基础设施。人工智能产业链一般为三层结构，包括基础层、技术层和应用层，其中基础层是人工智能产业的基础，为人工智能提供数据及算力支撑。服务器一般可分为通用服务器、云计算服务器、边缘服务器、AI 服务器等类型，AI 服务器专为人工智能训练和推理应用而设计。大模型兴起和生成式 AI 应用显著提升了对高性能计算资源的需求，AI 服务器是支撑这些复杂人工智能应用的核心基础设施，AI 服务器的其核心器件包括 CPU、GPU、FPGA、NPU、存储器等芯片，以及 PCB、高速连接器等。

图 11：人工智能系统产业链结构图



资料来源：电子工程世界，中原证券研究所

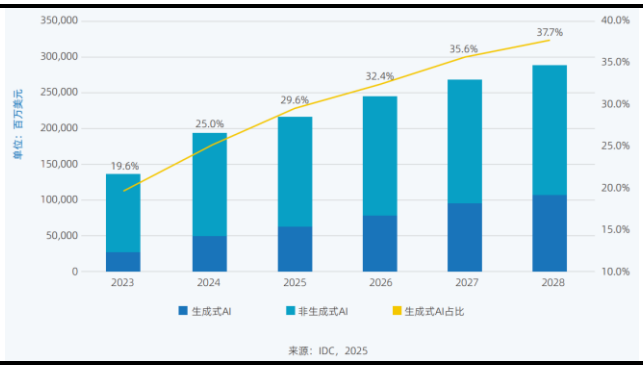
图 12：AI 服务器内部结构图



资料来源：McKinsey，中原证券研究所

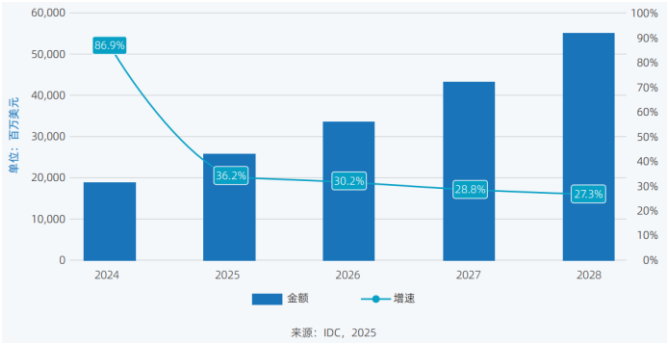
大模型有望推动 AI 服务器出货量高速成长。大模型带来算力的巨量需求，有望进一步推动 AI 服务器市场的增长。根据 IDC 的数据，2024 年全球 AI 服务器市场规模预计为 1251 亿美元，2025 年将增至 1587 亿美元，2028 年有望达到 2227 亿美元，2024-2028 年复合增速达 15.5%，其中生成式 AI 服务器占比将从 2025 年的 29.6%提升至 2028 年的 37.7%。IDC 预计 2024 年中国 AI 服务器市场规模为 190 亿美元，2025 年将达 259 亿美元，同比增长 36.2%，2028 年将达到 552 亿美元，2024-2028 年复合增速达 30.6%。

图 13：2023-2028 年全球生成式人工智能和非生成式人工智能服务器市场规模及预测



资料来源：IDC，2025 年中国人工智能计算力发展评估报告，中原证券研究所

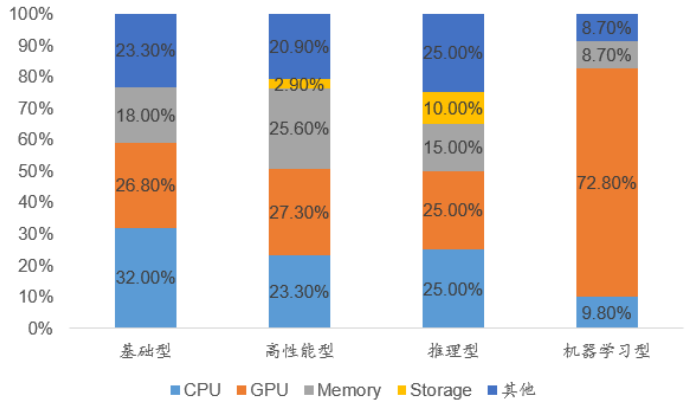
图 14：2024-2028 年中国 AI 服务器市场规模及预测



资料来源：IDC，2025 年中国人工智能计算力发展评估报告，中原证券研究所

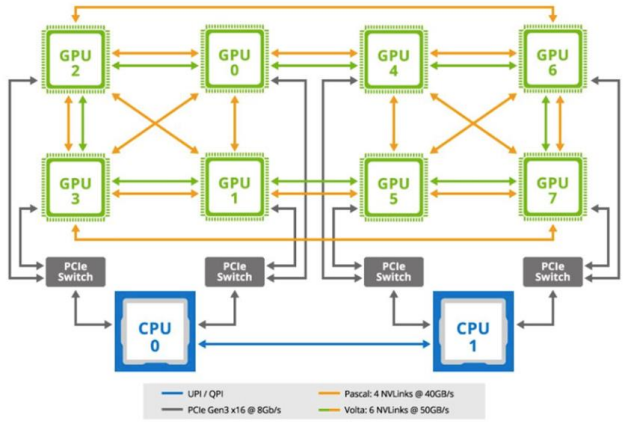
AI 算力芯片是算力的基石。CPU+GPU 是目前 AI 服务器主流的异构计算系统方案，根据 IDC 2018 年服务器成本构成的数据，推理型和机器学习型服务器中 CPU+GPU 成本占比达到 50-82.6%，其中机器学习型服务器 GPU 成本占比达到 72.8%。AI 算力芯片具备强大的并行计算能力，能够快速处理大规模数据和复杂的神经网络模型，并实现人工智能训练与推理任务；AI 算力芯片占 AI 服务器成本主要部分，为 AI 服务器提供算力的底层支撑，是算力的基石。AI 算力芯片作为“AI 时代的引擎”，有望畅享 AI 算力需求爆发浪潮，并推动 AI 技术的快速发展和广泛应用。

图 15：2018 年服务器成本构成情况



资料来源：IDC，智研咨询，中原证券

图 16：CPU+GPU 异构计算系统方案框图



资料来源：英伟达，中原证券

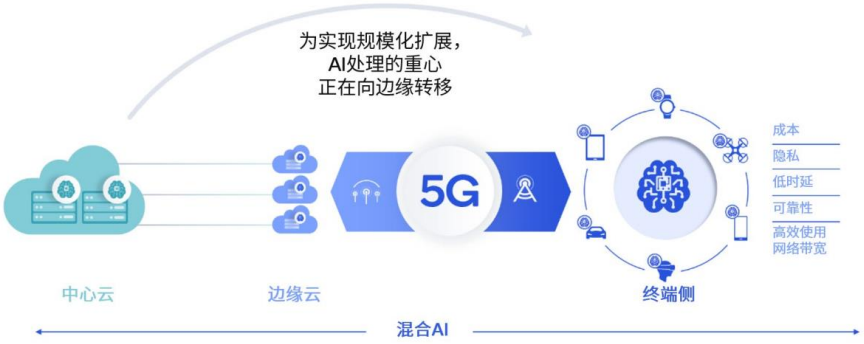
2. AI 算力芯片以 GPU 为主流，定制 ASIC 芯片市场高速成长

2.1. AI 算力芯片可应用于云端、边缘端、终端，当前以 GPU 为主流

混合 AI 是 AI 的发展趋势。AI 训练和推理受限于大型复杂模型而在云端部署，而 AI 推理的规模远高于 AI 训练，在云端进行推理的成本极高，将影响规模化扩展。随着生成式 AI 的快速发展以及计算需求的日益增长，AI 处理必须分布在云端和终端进行，才能实现 AI 的规模化扩展并发挥其最大潜能。混合 AI 指终端和云端协同工作，在适当的场景和时间下分配 AI 计算的工作负载，以提供更好的体验，并高效利用资源；在一些场景下，计算将主要以终端为中

心，在必要时向云端分流任务；而在以云为中心的场景下，终端将根据自身能力，在可能的情况下从云端分担一些 AI 工作负载。与仅在云端进行处理不同，混合 AI 架构在云端和边缘终端之间分配并协调 AI 工作负载；云端和边缘终端如智能手机、汽车、个人电脑和物联网终端协同工作，能够实现更强大、更高效且高度优化的 AI。

图 17：AI 处理的重心正在从云端向边缘转移



资料来源：高通，中原证券研究所

AI 算力芯片按应用场景可分为云端、边缘端、终端 AI 算力芯片。人工智能的各类应用场景，从云端溢出到边缘端，或下沉到终端，都需要由 AI 算力芯片提供计算能力支撑。云端、边缘端、终端三种场景对于 AI 算力芯片的运算能力和功耗等特性有着不同要求，云端 AI 算力芯片承载处理海量数据和计算任务，需要高性能、高计算密度，对于算力要求最高；终端对低功耗、高能效有更高要求，通常对算力要求相对偏低；边缘端对功耗、性能的要求通常介于终端与云端之间；本文主要针对于云端 AI 算力芯片。

表 1：云端、边缘端、终端应用场景对 AI 算力芯片的算力需求情况

应用场景	芯片需求	典型计算能力	典型功耗	典型应用领域
云端	高性能、高计算密度、兼有推理和训练任务、单价高、硬件产品形态少	>30TOPS	>50 瓦	云计算数据中心、企业私有云等
边缘端	对功耗、性能、尺寸的要求常介于终端与云端之间、推理任务为主、多用于插电设备、硬件产品形态相对较少	5TOPS 至 30TOPS	4 瓦至 15 瓦	智能制造、智能家居、智能零售、智慧交通、智慧金融、智慧医疗、智能驾驶等领域
终端	低功耗、高能效、推理任务为主、成本敏感、硬件产品形态众多	<8TOPS	<5 瓦	各类消费类电子、物联网产品

资料来源：寒武纪招股说明书，中原证券

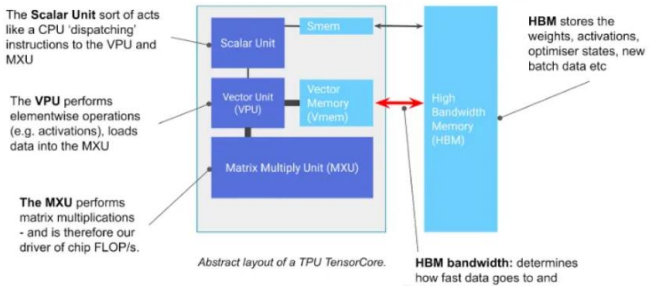
根据芯片的设计方法及应用，AI 算力芯片可分为通用型 AI 芯片和专用型 AI 芯片。通用型 AI 芯片为实现通用任务设计的芯片，主要包括 CPU、GPU、FPGA 等；专用型 AI 芯片是专门针对人工智能领域设计的芯片，主要包括 TPU（Tensor Processing Unit）、NPU（Neural Network Processing Unit）、ASIC 等。在通用型 AI 芯片中，由于在计算架构和性能特点上的不同，CPU 适合处理逻辑复杂、顺序性强的串行任务；GPU 是为图形渲染和并行计算设计的处理器，具有大量的计算核心，适合处理大规模并行任务；FPGA 通过集成大量的可重构逻辑单元阵列，可支持硬件架构的重构，从而灵活支持不同的人工智能模型。专用型 AI

芯片是针对面向特定的、具体的、相对单一的人工智能应用专门设计的芯片，其架构和指令集针对人工智能领域中的各类算法和应用作了专门优化，具体实现方法为在架构层面对特定智能算法作硬化支持，可高效支持视觉、语音、自然语言处理和传统机器学习等智能处理任务。

图 18：英伟达 A100 GPU 内部架构图

资料来源：英伟达，中原证券研究所

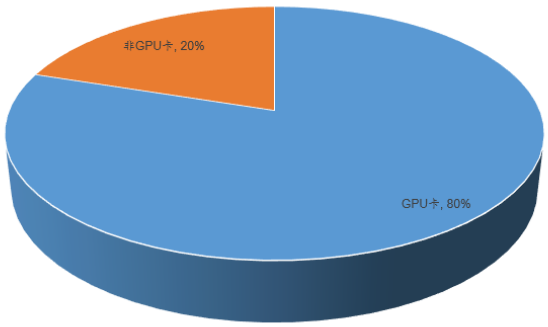
图 19：谷歌 TPU 内部架构图



资料来源：半导体行业观察，中原证券研究所

当前 AI 算力芯片以 GPU 为主流，英伟达主导全球 AI 算力芯片市场。根据的 IDC 数据，2024 上半年，中国 AI 加速芯片的市场规模达超过 90 万张；从技术角度来看，GPU 卡占据 80%的市场份额。根据 Precedence Research 数据，2022 年英伟达占据全球 AI 芯片市场份额超过 80%，其中英伟达占全球 AI 服务器加速芯片市场份额超过 95%。

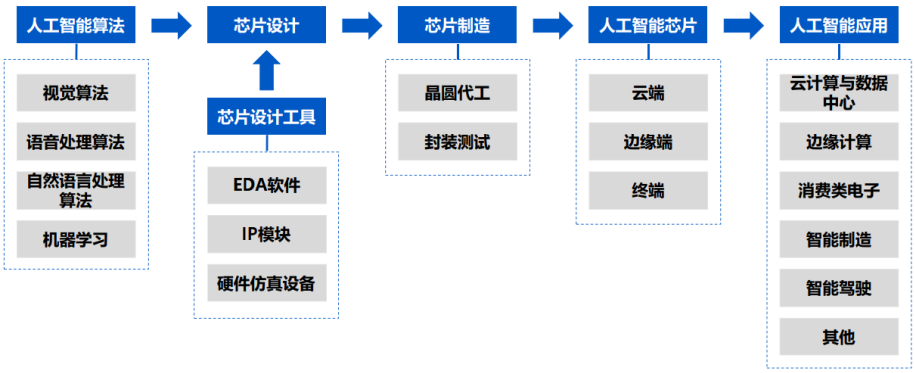
图 20：2024 年上半年中国 AI 芯片市场份额情况



资料来源：IDC，中原证券研究所

AI 算力芯片产业链包括人工智能算法、芯片设计、芯片制造及下游应用环节。人工智能芯片产业链上游主要是人工智能算法以及芯片设计工具，人工智能算法覆盖广泛，包括视觉算法、语音处理算法、自然语言处理算法以及各类机器学习方法（如深度学习等）。AI 算力芯片行业的核心为芯片设计和芯片制造，芯片设计工具厂商、晶圆代工厂商与封装测试厂商为 AI 算力芯片提供了研发工具和产业支撑。 AI 算力芯片行业的下游应用场景主要包括云计算与数据中心、边缘计算、消费类电子、智能制造、智能驾驶、 智慧金融、智能教育等领域。

图 21：AI 算力芯片产业链结构图



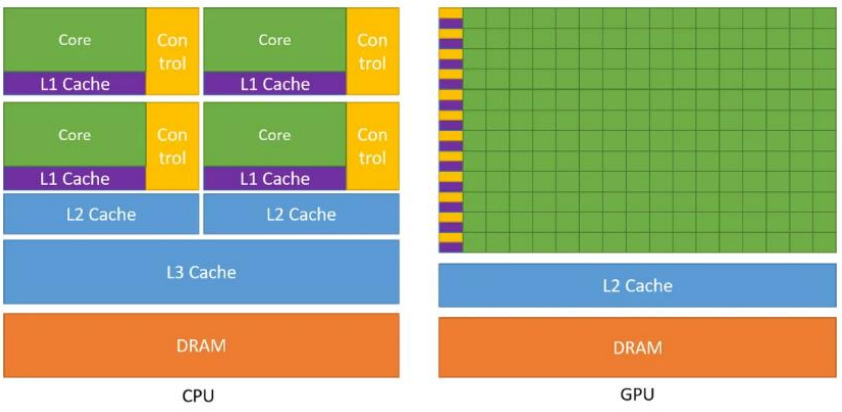
资料来源：寒武纪招股说明书，中原证券研究所

2.2. 英伟达主导全球 GPU 市场，GPU 生态体系建立极高的行业壁垒

GPU（Graphics Processing Unit）即图形处理单元，是计算机的图形处理及并行计算的核心。GPU 最初主要应用于加速图形渲染，如 3D 渲染、图像处理和视频解码等，是计算机显卡的核心；随着技术的发展，GPU 也被广泛应用于通用计算领域，如人工智能、深度学习、科学计算、大数据处理等领域，用于通用计算的 GPU 被称为 GPGPU（General-Purpose computing on Graphics Processing Units），即通用 GPU。

GPU 与 CPU 在内部架构上有显著差异，决定了它们各自的优势领域。GPU 通过大量简单核心和高带宽内存架构，优化并行计算能力，适合处理大规模数据和高吞吐量任务；CPU 通过少量高性能核心和复杂控制单元优化单线程性能，适合复杂任务和低延迟需求。

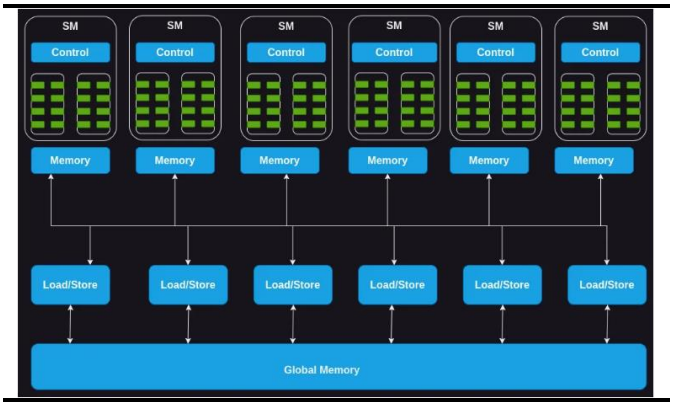
图 22：GPU 与 CPU 内部架构对比图



资料来源：英伟达，OneFlow，中原证券研究所

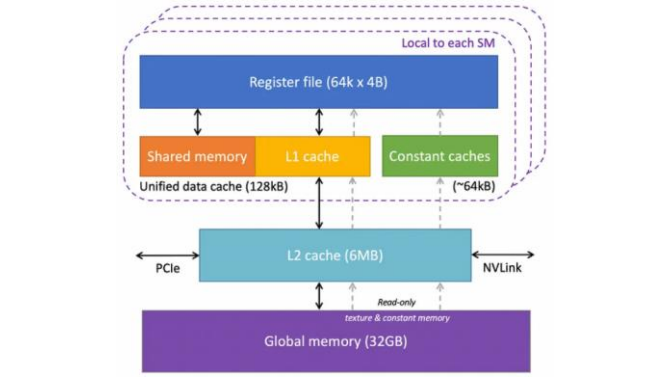
GPU 架构由流处理器（SM）、光栅操作单元、纹理单元、专用加速单元等多个关键组件组成，这些组件协同工作，以实现高效的通用计算和图形渲染。GPU 的计算架构由一系列流式多处理器（SM）组成，其中每个 SM 又由多个流式处理器、核心或线程组成，例如，NVIDIA H100 GPU 具有 132 个 SM，每个 SM 拥有 64 个核心，总计核心高达 8448 个；每个 SM 还配备了几个功能单元或其他加速计算单元，例如张量核心（Tensor Core）或光线追踪单元（Ray Tracing Unit），用于满足 GPU 所处理的工作负载的特定计算需求。GPU 具有多层不同类型的内存，每一层都有其特定用途。

图 23：GPU 的计算架构



资料来源：OneFlow，中原证券研究所

图 24：GPU 的内存架构



资料来源：OneFlow，中原证券研究所

GPU 硬件性能可以通过多个参数综合评估，包括核心数量、核心频率、显存容量、显存位宽、显存带宽、显存频率、工艺制程等。GPU 的核心数量越多、核心频率越高，GPU 的计算能力越强。显存容量越大，GPU 能够处理的数据规模就越大；显存带宽越高，GPU 显存与核心之间数据传输的速率越快。GPU 的工艺制程越先进，GPU 性能越好、功耗越低。

表 2：GPU 硬件性能评价参数

性能参数	含义
CUDA 核心数量	CUDA 核心是英伟达 GPU 中用于进行通用计算的处理单元，数量越多，GPU 并行处理数据的能力就越强。
Tensor 核心数量	Tensor 核心是英伟达 GPU 中的专用硬件单元，主要用于加速 AI 和深度学习任务；Tensor 核心数量越多性能越好，Tensor 核心的性能随着架构升级而不断提升；Tensor 核心的性能优势可以通过高吞吐量、混合精度支持及性能等方面来体现。
核心频率	核心频率是指 GPU 每秒钟执行的次数，核心频率越高，性能越强。
显存容量	显存容量决定着显存临时存储数据的多少，显存容量越大，GPU 能够处理的数据规模就越大。
显存带宽	显存带宽是指 GPU 显存与核心之间数据传输的速率，它反映了 GPU 在单位时间内能够处理的数据量。显存带宽＝显存频率×显存位宽/8，显存带宽与显存频率、显存位宽成正比关系。
显存位宽	显存位宽是指 GPU 显存接口的数据传输通道的宽度，通常以 bit（位）为单位。显存位宽越大，GPU 与显存之间每次可以传输的数据量越多，显存带宽越高。
显存频率	显存频率是指显存在单位时间内能够进行数据传输的次数，通常以 MHz 为单位，显存频率决定了显存与 GPU 之间数据传输的速度。
工艺制程	工艺制程是指在制造 GPU 芯片时所采用的技术工艺和制造流程，通常用纳米（nm）来衡量，工艺制程越先进，GPU 性能越好、功耗越低。

资料来源：平行云，华秋商城，中原证券研究所

GPU 架构对性能影响至关重要，不同架构下的硬件性能参数有所不同。GPU 架构的每次升级在计算能力、图形处理能力、能效比等多方面对性能产生了显著提升，所以 GPU 架构对性能影响至关重要。通过对比英伟达 GeForce 系列 RTX 3090、RTX 4090、RTX 5090，不同 GPU 架构下硬件性能参数有所不同。随着 GPU 架构的升级，GPU 厂商通常会采用更先进的工艺制程，比如英伟达从 8nm 工艺的 Ampere 架构升级到 4nm 工艺的 Blackwell 架构，在相同性能下，新工艺能够降低功耗，或者在相同功耗下提供更高的性能。

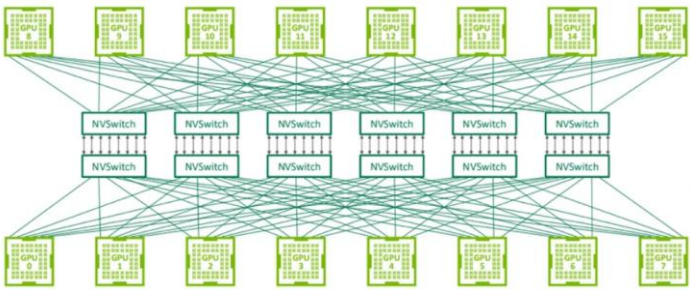
表 3：英伟达 GeForce 系列 GPU 硬件性能参数对比情况

	RTX 3090	RTX 4090	RTX 5090
GPU 架构	NVIDIA Ampere	NVIDIA Ada Lovelac	NVIDIA Blackwell
CUDA 核心数量	10496	16384	21760
Tensor 核心数量	328	512	680
核心频率	1.70 GHz	2.52 GHz	2.41 GHz
显存容量	24 GB	24 GB	32 GB
显存带宽	936 GB/s	1008 GB/s	1792 GB/s
显存位宽	384 bit	384 bit	512 bit
显存频率	19.5 Gbps	21 Gbps	28 Gbps
工艺制程	Samsung 8 nm 8N	TSMC 4nm 4N	TSMC 4nm 4N

资料来源：英伟达，中原证券研究所

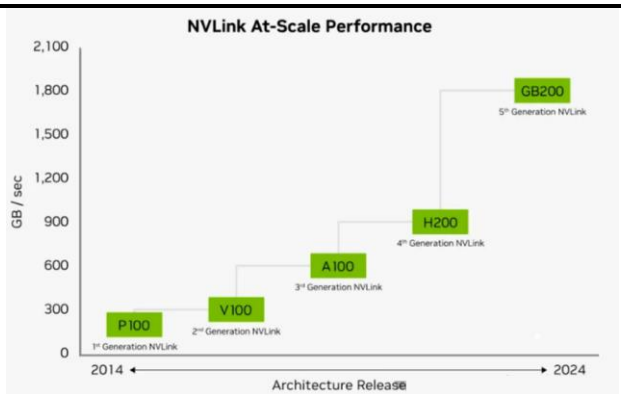
多 GPU 互连成为行业发展趋势，以提高系统的计算能力。随着 AI 大模型时代来临，AI 算力需求不断增长，由于单 GPU 芯片算力和内存有限，无法承载大模型的训练任务，通过多种互连技术将多颗 GPU 芯片互连在一起提供大规模的算力，已成为行业发展趋势。对于多 GPU 系统，如何实现 GPU 之间的高速数据传输和协同工作是关键问题。英伟达推出 NVLink、NVSwitch 等互连技术，通过更高的带宽和更低的延迟，为多 GPU 系统提供更高的性能和效率，支持 GPU 之间的高速数据传输和协同工作，提高通信速度，加速计算过程等。NVLink 用于连接多个 GPU 之间或连接 GPU 与其他设备（如 CPU、内存等）之间的通信，它允许 GPU 之间以点对点方式进行通信，具有比传统的 PCIe 总线更高的带宽和更低的延迟。NVSwitch 实现单服务器中多个 GPU 之间的全连接，允许单个服务器节点中多达 16 个 GPU 实现全互联，每个 GPU 都可以与其他 GPU 直接通信，无需通过 CPU 或其他中介。经过多年演进，NVLink 技术已升级到第 5 代，NVLink 5.0 数据传输速率达到 100GB/s，每个 Blackwell GPU 有 18 个 NVLink 连接，Blackwell GPU 将提供 1.8TB/s 的总带宽，是 PCIe Gen5 总线带宽的 14 倍；NVSwitch 也升级到了第四代，每个 NVSwitch 支持 144 个 NVLink 端口，无阻塞交换容量为 14.4TB/s。

图 25：英伟达多 GPU 系统架构图



资料来源：nextplatform，半导体行业观察，中原证券研究所

图 26：英伟达 NVLink 技术演进情况



资料来源：英伟达，半导体行业观察，中原证券研究所

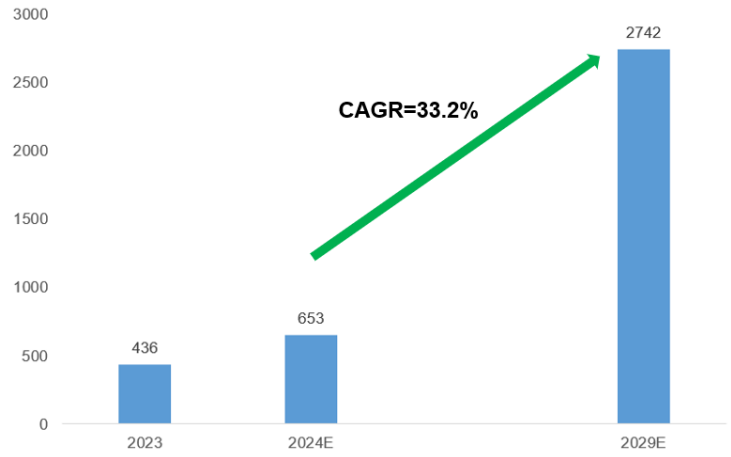
GPU 应用场景广泛，数据中心 GPU 市场快速增长。GPU 最初设计用于图形渲染，但随着其并行计算能力的提升，GPU 的应用场景已经扩展到数据中心、自动驾驶、机器人、区块链与加密货币、科学计算、金融科技、医疗健康等多个领域。近年来数据中心 GPU 市场在全球范围内呈现出快速增长的趋势，尤其是在人工智能、高性能计算和云计算等领域。

图 27：GPU 应用场景广泛

资料来源：极云科技，中国算力发展报告（2024 年），中原证券研究所

GPU 是 AI 服务器算力的基石，有望畅享 AI 算力需求爆发浪潮。GPU 是 AI 服务器算力的基石，随着 AI 算力规模的快速增长将催生更大的 GPU 芯片需求。根据 Statista 的数据，2023 年全球 GPU 市场规模为 436 亿美元，预计 2029 年市场规模将达到 2742 亿美元，预计 2024-2029 年复合增速达 33.2%。

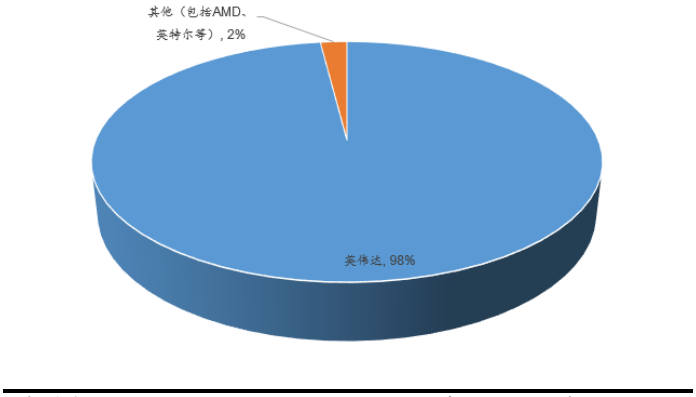
图 28：2023-2029 全球 GPU 市场规模情况及预测（亿美元）



资料来源：Statista，半导体行业观察，中原证券研究所

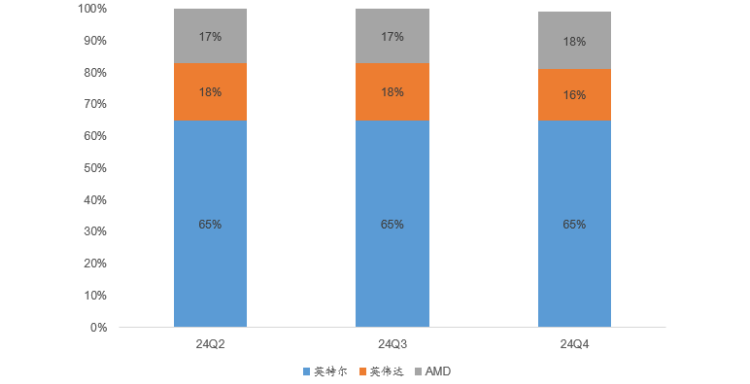
英伟达主导全球 GPU 市场。根据 TechInsights 的数据，2023 年全球数据中心 GPU 总出货量达到了 385 万颗，相比 2022 年的 267 万颗同比增长 44.2%，其中英伟达数据中心 2023 年 GPU 出货量呈现爆发式增长，总计约 376 万台，英伟达在数据中心 GPU 出货量中占据 98%的市场份额，英伟达还占据全球数据中心 GPU 市场 98% 的收入份额，达到 362 亿美元，是 2022 年 109 亿美元的三倍多。根据 Jon Peddie Research 的数据， 2024 年第四季度全球 PC GPU 出货量达到 7800 万颗，同比增长 0.8%，环比增长 6.2%，其中英特尔、AMD、英伟达的市场份额分别为 65%、18%、16%。

图 29：2023 年全球数据中心 GPU 市场竞争格局情况



资料来源：TechInsights，半导体行业观察，中原证券研究所

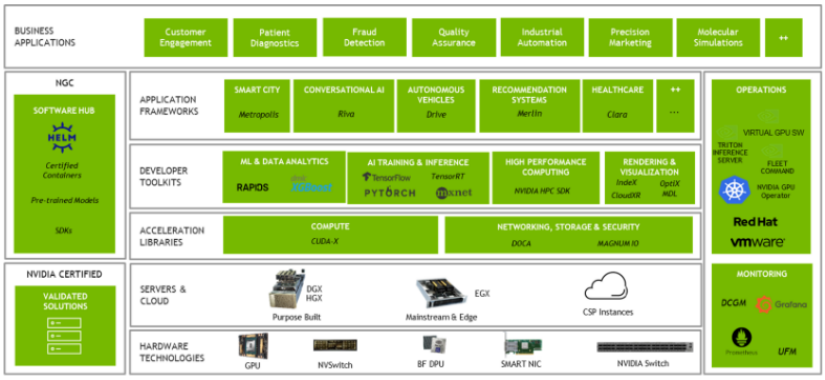
图 30：24Q4 全球 PC GPU 市场竞争格局情况



资料来源：Jon Peddie Research，快科技，中原证券研究所

GPU 生态体系主要由三部分构成，包括底层硬件，中间层 API 接口、算法库、开发工具等，上层应用。以英伟达数据中心平台 GPU 生态体系为例，底层硬件的核心是英伟达的 GPU 产品、用于 GPU 之间高速连接的 NVSwitch 、节点之间互联的各种高速网卡、交换机等，以及基于 GPU 构建的服务器；中间层是软件层面的建设，包括计算相关的 CUDA-X、网络存储及安全相关的 DOCA 和 MAGNUM IO 加速库，以及编译器、调试和优化工具等开发者工具包和基于各种行业的应用框架；上层是开发者基于英伟达提供的软硬件平台能力，所构建的行业应用。

图 31：英伟达数据中心平台 GPU 生态体系架构图

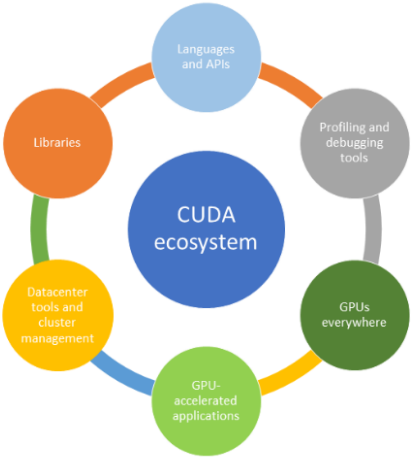


资料来源：英伟达，infoQ，中原证券研究所

GPU 厂商非常重视软件生态系统的构建，英伟达 CUDA 生态几乎占据通用计算 GPU 领域的全部市场。CUDA 全称为 Compute Unified Device Architecture，即统一计算设备架构，是英伟达推出的基于其 GPU 的通用高性能计算平台和编程模型。目前 CUDA 生态包括编程语言和 API、开发库、分析和调试工具、GPU 加速应用程序、GPU 与 CUDA 架构链接、数据中心工具和集群管理六个部分。编程语言和 API 支持 C、C++、Fortran、Python 等多种高级编程语言；英伟达提供的 CUDA 工具包可用于在 GPU 上开发、优化和部署应用程序，还支持第三方工具链，如 PyCUDA、AltiMesh Hybridizer、OpenACC、OpenCL、Alea - GPU 等，方便开发者从不同的编程接口来使用 CUDA。英伟达在 CUDA 平台上提供了 CUDA-X，它是一系列库、工具和技术的集合，其中包括数学库、并行算法库、图像和视频库、通信库、深度学习库等，同时还支持 OpenCV、FFmpeg 等合作伙伴提供的库。英伟达提供了多种工具来帮助开发者进行性能分析和调试， NVIDIA Nsight 是低开销的性能分析、跟踪和调试工具，提

供基于图形用户界面的环境，可在多种英伟达平台上使用；CUDA GDB 是 Linux GDB 的扩展，提供基于控制台的调试接口；CUDA - Memcheck 可用于检查内存访问问题；此外还支持第三方解决方案，如 ARM Forge、TotalView Debugger 等。目前几乎所有的深度学习框架都使用 CUDA/GPU 计算来加速深度学习的训练和推理，英伟达维护了大量经过 GPU 加速的应用程序。在数据中心中，英伟达与生态系统合作伙伴紧密合作，为开发者和运维人员提供软件工具，涵盖 AI 和高性能计算软件生命周期的各个环节，以实现数据中心的轻松部署、管理和运行；例如通过 Mellanox 高速互连技术，可将数千个 GPU 连接起来，构建大规模的计算集群。CUDA 生态系统复杂，建设难度大，CUDA 生态几乎占据通用计算 GPU 领域的全部市场。

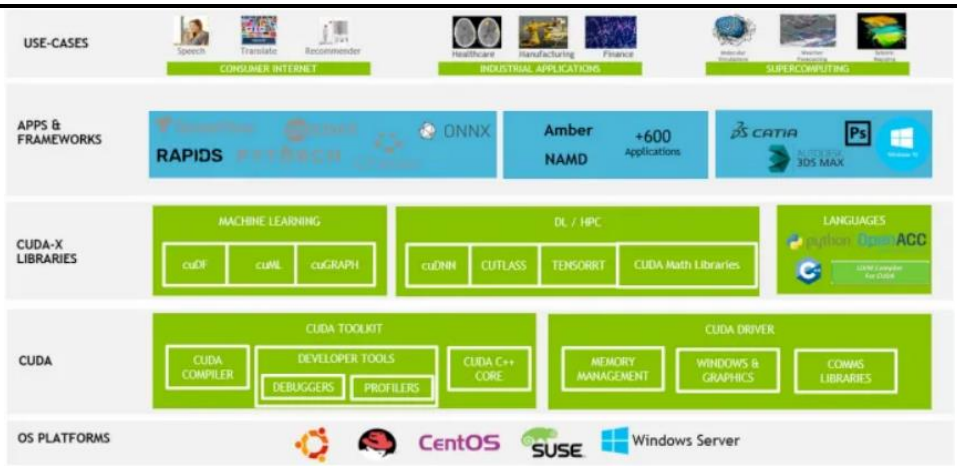
图 32：英伟达 CUDA 生态系统的组成



资料来源：英伟达，中原证券研究所

GPU 生态体系建立极高的行业壁垒。GPU 一方面有对硬件性能的要求，还需要软件体系进行配套，而 GPU 软件生态系统复杂，建设周期长、难度大。英伟达 CUDA 生态从 2006 年开始建设，经过多年的积累，建立强大的先发优势，英伟达通过与客户进行平台适配、软件开源合作，不断加强客户粘性，GPU 行业新进入者转移客户的难度极大，GPU 生态体系建立极高的行业壁垒。

图 33：英伟达 CUDA 加速计算解决方案

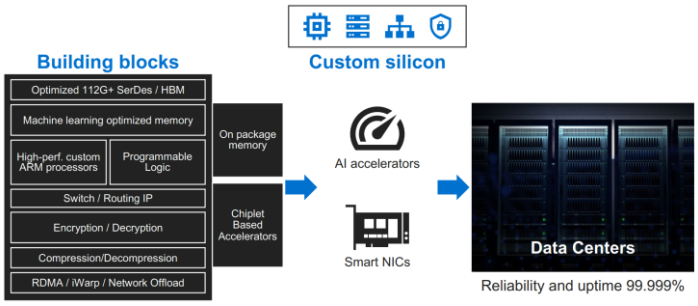


资料来源：英伟达，半导体行业研究，中原证券研究所

2.3. 云厂商推动定制 ASIC 芯片市场高速成长

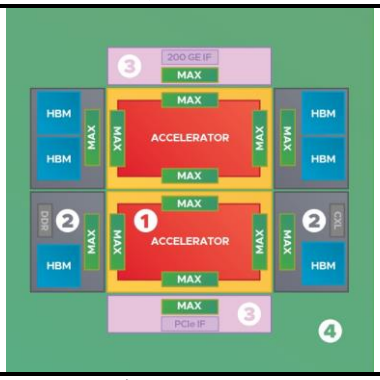
AI ASIC 是一种专为人工智能应用设计的定制集成电路，具有高性能、低功耗、定制化、低成本等特点。与通用处理器相比，AI ASIC 针对特定的 AI 任务和算法进行了优化，如深度学习中的矩阵乘法、卷积等运算，能在短时间内完成大量计算任务，提供高吞吐量和低延迟，满足 AI 应用对实时性的要求；AI ASIC 通过优化电路设计和采用先进的工艺技术，减少不必要的能耗，在处理 AI 工作负载时具有较高的能效比，适合大规模数据中心等对能耗敏感的场景；虽然前期研发和设计成本较高，在大规模部署时，ASIC 的单位计算成本通常低于通用处理器。

图 34：Marvell 用于数据中心的 ASIC 解决方案



资料来源：Marvell，中原证券研究所

图 35：博通 AI ASIC 内部架构图



资料来源：博通，中原证券研究所

AI ASIC 与 GPU 在 AI 计算任务中各有优势和劣势。在算力上，先进 GPU 比 ASIC 有明显的优势；ASIC 针对特定任务优化，通常能提供更高的计算效率，ASIC 在矩阵乘法、卷积运算等特定 AI 任务上性能可能优于 GPU；GPU 通用性强，能够运行各种不同类型的算法和模型，ASIC 功能固定，难以修改和扩展，灵活性较差；ASIC 针对特定任务优化，功耗显著低于 GPU；GPU 研发和制造成本较高，硬件成本是大规模部署的重要制约因素，ASIC 在大规模量产时单位成本相对较低。

表 4：AI ASIC 与 GPU 性能参数对比情况

厂商	产品型号	发布时间	工艺	核心数量	FP32 算力	TF32 算力	FP/BF16 算力	INT8 算力	显存容量	显存带宽	芯片间互联带宽	功耗
			nm		TFLOPS	TFLOPS	TFLOPS	TOPS	GB	GB/s	GB/s	W
英伟达	H100 SXM	2022	4	16896	67	989	1979	3958	80	3350	900	700
英伟达	GB200	2024	4	20480	180	5000	10000	20000	384	16000	3600	
AMD	MI250X	2021	6	14080	95.7		383	383	128	3200	800	560
AMD	MI300X	2023	5/6	19456	163.4	653.7	1307.4	2614.9	192	5300	896	750
谷歌	TPU v5p	2023	5				459	918			1200	
谷歌	TPU v6 Trillium	2024	4				926	1852				
亚马逊	Trainium 2	2023			181		667				1280	
Meta	MTIA v2	2024	5				354	708				90
微软	Maia 100	2024	5				800	1600				700

资料来源：各公司官网，STH，The Next Platform，中原证券研究所

GPU 软件生态成熟且丰富，AI ASIC 推动软件生态走向多元化。ASIC 的软件生态缺乏通用性，主要是对特定应用场景和算法进行优化；由于 ASIC 的开发工具和软件库资源相对较

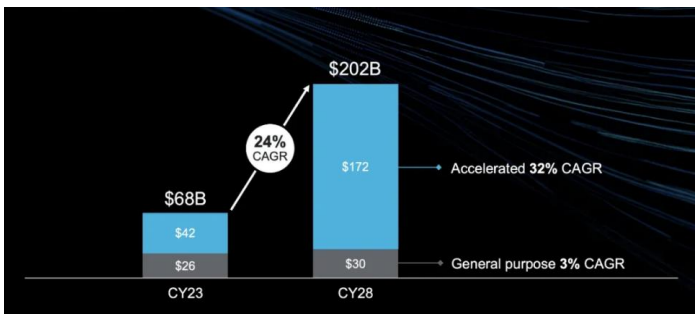
少，编程难度比 GPU 大，开发者在使用 ASIC 进行开发和调试时所需要花费时间会更多。GPU 的软件生态成熟且丰富，如英伟达 CUDA 和 AMD ROCm 等，提供了广泛的开发工具、编程语言支持，并拥有大量的开源项目和社区资源。为了提升 AI ASIC 在特定场景下的计算效率，谷歌、亚马逊、META、微软等厂商为 ASIC 开发了配套的全栈软件生态，包括编译器、底层中间件等，持续降低从 CUDA 生态向其他生态转换的迁移成本，以减轻对 CUDA 生态的依赖性。

图 36：华为昇腾 AI 生态系统架构图

资料来源：华为，搜狐，中原证券研究所

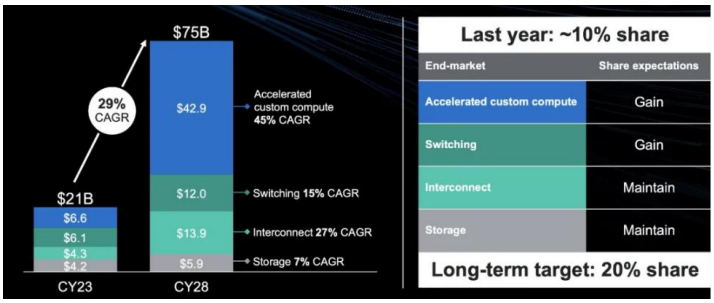
云厂商推动数据中心定制 ASIC 芯片市场高速增长，预计增速快于通用 AI 算力芯片。由于全球头部云厂商对 AI 算力芯片需求量巨大，英伟达垄断全球数据中心 GPU 市场，过度依赖单一供应商风险较大，为了提升议价能力及供应链多元化，云厂商大力投入自研 AI ASIC，推动数据中心定制 ASIC 芯片市场高速增长。根据 Marvell 的数据，2023 年数据中心 AI 算力芯片市场规模约为 420 亿美元，其中定制 ASIC 芯片占比 16%，市场规模约为 66 亿美元；预计 2028 年数据中心定制 ASIC 芯片市场规模将达到 429 亿美元，市场份额约为 25%，2023-2028 年复合增速将达到 45%；预计 2028 年数据中心 AI 算力芯片市场规模将达约 1720 亿美元，2023-2028 年复合增速约为 32%。

图 37：2023-2028 年数据中心 AI 算力芯片市场规模及预测情况



资料来源：650 Group, CignalAI, Dell'Oro, LightCounting, Marvell，半导体行业观察，中原证券研究所

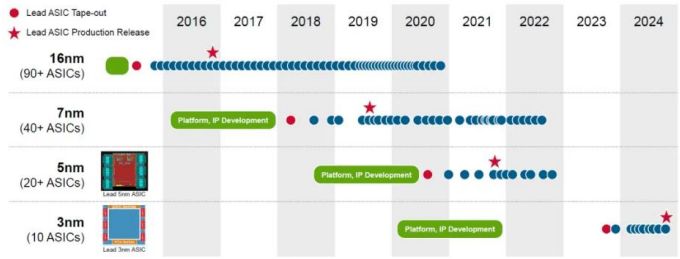
图 38：2023-2028 年数据中心 ASIC 定制芯片市场规模及预测情况



资料来源：650 Group, CignalAI, Dell'Oro, LightCounting, Marvell，半导体行业观察，中原证券研究所

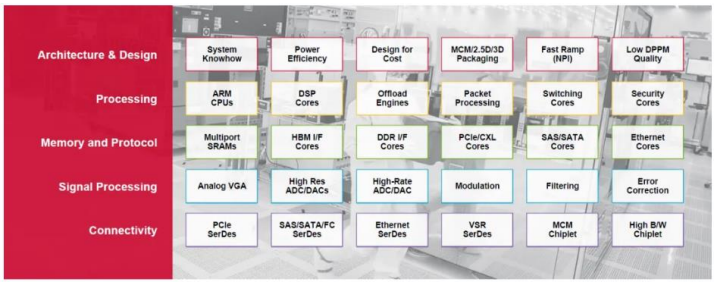
云厂商自研 AI ASIC 芯片时，通常会与芯片设计厂商合作，然后再由台积电等晶圆代工厂进行芯片制造，目前全球定制 AI ASIC 市场竞争格局以博通、Marvell 等厂商为主。博通为全球定制 AI ASIC 市场领导厂商，已经为大客户实现 AI ASIC 大规模量产。博通在多年的发展中已经积累了大量的成体系的高性能计算/互连 IP 核及相关技术，除了传统的 CPU/DSP IP 核外，博通还具有交换、互连接口、存储接口等关键 IP 核；这些成体系的 IP 核可以帮助博通降低 ASIC 产品成本和研发周期，以及降低不同 IP 核联合使用的设计风险，并建立博通强大的竞争优势。博通 2024 财年 AI 收入达到 120 亿美元，公司 CEO 表示，到 2027 年，公司在 AI 芯片和网络组件的市场规模将达到 600 亿到 900 亿美元。

图 39：博通累积的定制芯片设计经历



资料来源：博通，半导体产业纵横，中原证券研究所

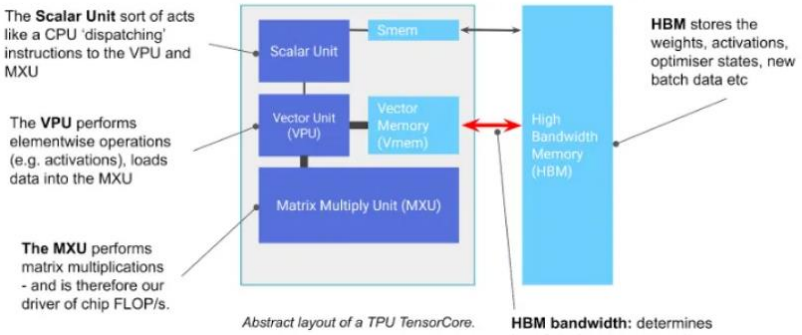
图 40：博通定制技术能力与 IP 核情况



资料来源：博通，半导体产业纵横，中原证券研究所

谷歌 TPU（Tensor Processing Unit）即张量处理单元，是谷歌专为加速机器学习任务设计的定制 ASIC 芯片，主要用于深度学习的训练和推理。TPU 基本上是专门用于矩阵乘法的计算核心，并与高带宽内存（HBM）连接；TPU 的基本组件包括矩阵乘法单元 (MXU)、矢量单元 (VPU) 和矢量内存 (VMEM)；矩阵乘法单元是 TensorCore 的核心，矢量处理单元执行一般数学运算，矢量内存是位于 TensorCore 中靠近计算单元的片上暂存器；TPU 在进行矩阵乘法方面速度非常快。

图 41：TPU 内部架构图



资料来源：半导体行业观察，中原证券研究所

目前谷歌 TPU 已经迭代至第六代产品，每代产品相较于上一代在芯片架构及性能上均有一定的提升。2015 年谷歌 TPU v1 推出，主要用于推理任务。2024 年谷歌发布第六代产品 TPU v6 Trillium，是目前性能最强、能效最高的 TPU。TPU v6 Trillium 与上一代 TPU v5e 相比，单芯片峰值计算性能提高了 4.7 倍，HBM 容量和带宽均增加一倍，同时芯片间互连带宽也增加一倍；TPU v6 Trillium 在性能提升的同时，能源效率比上一代提高了 67%，显著降低

了运营成本；TPU v6 Trillium 被用于训练谷歌的 Gemini 2.0 等 AI 大模型。

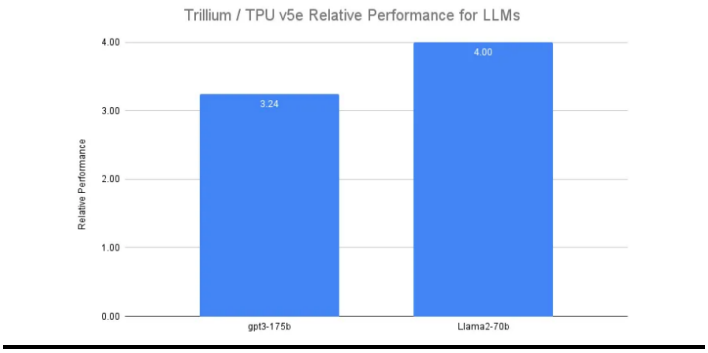
表 5：谷歌 TPU 历代产品性能参数情况

	v1	v2	v3	v4	v5e	v5p	v6 Trillium
发布时间	2015	2017	2018	2021	2023	2023	2024
BF16 算力（TFLOPs）	-	46	123	137.5	197	459	926
INT8 算力（TFLOPs）	92	-	-	275	394	918	1852
HBM 容量（GB）	8	16	32	32	16	95	32
HBM 带宽（GB/s）	300	700	900	1228	819	2765	1640
ICI 带宽（GB/s）	-	4*496	4*656	6*448	4*400	6*800	4*800
工艺制程（nm）	28	16	16	7	5	5	4

资料来源：Next Platform，中原证券研究所

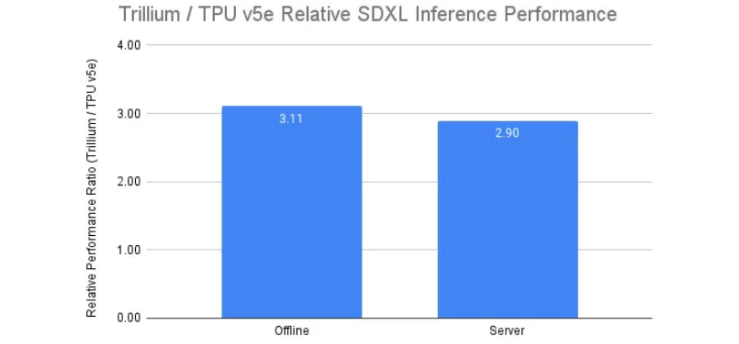
谷歌 TPU 迭代推动大模型训练与推理效率大幅提升。Gemini 等 AI 大模型性能强大且复杂，拥有数十亿个参数，训练如此密集的大模型 需要巨大的计算能力以及共同设计的软件优化。与上一代 TPU v5e 相比，TPU v6 Trillium 为 Llama-2-70b 和 gpt3-175b 等大模型提供了高达 4 倍的训练速度。TPU v6 Trillium 为推理工作负载提供了重大改进，为图像扩散和大模型 提供了最好的 TPU 推理性能，从而实现了更快、更高效的 AI 模型部署；与 TPU v5e 相比，TPU v6 Trillium 的 Stable Diffusion XL 离线推理相对吞吐量（每秒图像数）高出 3.1 倍，服务器推理相对吞吐量高出 2.9 倍。

图 42：在 TPU v5e 和 v6 Trillium 上运行的 steptime 的 Google 基准测试情况



资料来源：谷歌，半导体行业观察，中原证券研究所

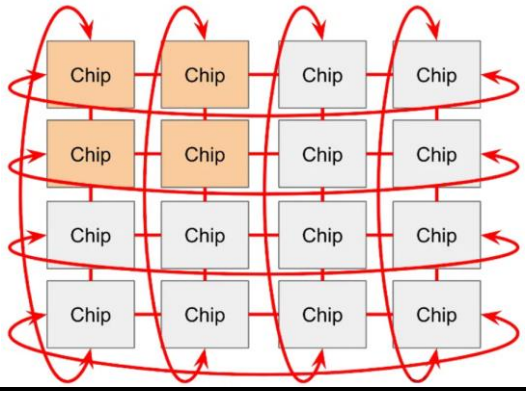
图 43：在 TPU v5e 和 v6 Trillium 上进行 SDXL 基准测试情况



资料来源：谷歌，半导体行业观察，中原证券研究所

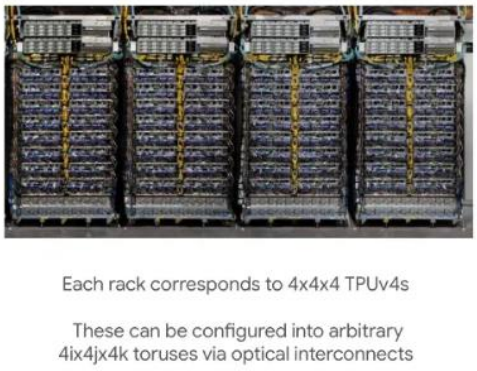
谷歌已建立 100000 TPU 芯片算力集群。TPU 芯片通过 ICI 连接成算力集群，TPU 网络可以连接 16x16x16 TPU v4 和 16x20x28TPU v5p。为了满足日益增长的 AI 计算需求，谷歌已将超过 100000 个 TPU v6 Trillium 芯片连接到一个网络结构中，构建了世界上最强大的 AI 超级计算机之一；该系统将超过 100000 个 TPU v6 Trillium 芯片与每秒 13 PB 带宽的 Jupiter 网络结构相结合，使单个分布式训练作业能够扩展到数十万个加速器上。这种大规模芯片集群可以提供强大的计算能力，实现高效的并行计算，从而加速大模型的训练过程，提高人工智能系统的性能和效率。

图 44：谷歌 TPU 芯片通过 ICI 相互连接



资料来源：半导体行业观察，中原证券研究所

图 45：由 TPU v4 建立的算力集群示意图



资料来源：半导体行业观察，中原证券研究所

2.4. 美国不断加大对高端 AI 算力芯片出口管制，国产厂商迎来黄金发展期

美国对高端 GPU 供应限制不断趋严，国产 AI 算力芯片厂商迎来黄金发展期。美国商务部在 2022、2023、2025 年连续对高端 AI 算力芯片进行出口管制，不断加大英伟达及 AMD 高端 GPU 芯片供应限制，国产 AI 算力芯片厂商迎来黄金发展机遇，但国产厂商华为海思、寒武纪、海光信息、壁仞科技和摩尔线程等进入出口管制“实体清单”，晶圆代工产能供应受限，影响国产 AI 算力芯片发展速度。

表 6：近年美国对 AI 算力芯片相关制裁政策情况

时间	具体事件及制裁政策情况
2022 年 8 月	美国芯片厂商英伟达和 AMD 收到美国政府通知，要求停止向中国出口用于人工智能的高端计算芯片，该禁令影响的芯片分别为英伟达的 GPU A100 与 H100，以及 AMD 的 GPU MI200。
2022 年 10 月	美国商务部公布一系列针对中国的出口管制新规，BIS 这项新的半导体出口限制政策涉及到对中国的先进计算、半导体先进制造进行出口管制；具体要限制美国的半导体设备在国内应用到 16/14nm 及以下工艺节点（非平面架构）的逻辑电路制造、128 层及以上的 3D NAND 工艺制造、18nm 及以下的 DRAM 工艺制造；对中国超级计算机或半导体开发或生产最终用途的项目进行限制；限制美国公民支持中国半导体制造或者研发。
2022 年 12 月	美国商务部将长江存储、上海微电子、寒武纪等 36 家中国实体加入出口管制“实体清单”。
2023 年 10 月	美国商务部公布针对先进计算芯片、半导体制造设备出口管制的更新规则，并将 13 家中国 GPU 企业列入实体清单，主要为壁仞科技和摩尔线程及其子公司。
2025 年 1 月	美国政府公布对 AI 芯片出口的新限制措施，这份新规将出口目的地分为三类，美国对 18 个关键盟友与合作伙伴的芯片销售无任何限制；对中国、伊朗等实施了严格的 AI 芯片销售限制；对其他国家，大多数国家则将面临总算力限制，每个国家在 2025 年至 2027 年期间最多可获得约 5 万个 AI GPU。
2025 年 1 月	美国商务部修订了《出口管制条例》，共增加了 25 个中国实体，主要包括智谱旗下 10 个实体、算能旗下约 11 个实体，以及哈勃投资的光刻机企业科益虹源等；BIS 还更新先进计算半导体的出口管制，针对于先进逻辑集成电路是采用“16nm/14nm 节点”及以下工艺、或采用非平面晶体管架构生产的逻辑集成电路，采取更多审查和规范，并且细化了多个物项信息如 DRAM 行业 18 纳米半间距节点的生产标准等。

资料来源：中华人民共和国商务部官网，美国商务部官网，美国政府官网，人民网，央视网，芯智讯，半导体产业纵横，腾讯，新浪，中原证券研究所

国产 AI 算力芯片厂商不断追赶海外龙头厂商，但在硬件性能上与全球领先水平仍有一定的差距。随着 AI 应用计算量的不断增加，要实现 AI 算力的持续大幅增长，既要单卡性能提升，又要多卡组合。从 AI 算力芯片硬件来看，单个芯片硬件性能及卡间互联性能是评估 AI 算力芯片产品水平的核心指标。国产厂商在芯片微架构、制程等方面不断追赶海外龙头厂商，产品性能逐步提升，但与全球领先水平仍有 1-2 代的差距。

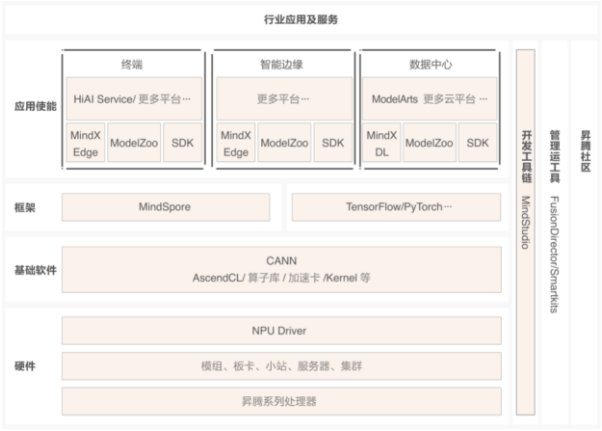
表 7：部分国产 AI 算力芯片技术指标与国际主流产品对比情况

厂商	产品型号	发布时间	工艺	核心数量	FP32 算力	TF32 算力	FP/BF16 算力	INT8 算力	显存容量	显存带宽	GPU 间互联带宽	功耗
			nm		TFLOPS	TFLOPS	TFLOPS	TOPS	GB	GB/s	GB/s	W
英伟达	V100 SXM	2017	12	5120	15.7	125			32	900	300	300
英伟达	A100 SXM	2020	7	6912	19.5	156	312	624	80	2039	600	400
英伟达	H100 SXM	2022	4	16896	67	989	1979	3958	80	3350	900	700
英伟达	GB200	2024	4	20480	180	5000	10000	20000	384	16000	3600	
AMD	MI100	2020	7	7680	23.1	46.1	92.3	92.3	32	1200	276	300
AMD	MI250X	2021	6	14080	95.7		383	383	128	3200	800	560
AMD	MI300X	2023	5/6	19456	163.4	653.7	1307.4	2614.9	192	5300	896	750
寒武纪	MLU370-X8	2022	7		24		96	256	48	614.4	200	250
海光信息	深算一号	2021	7	4096					32	1024	184	350
华为	昇腾 910	2019	7				256	512				
壁仞科技	壁砺™ 106B	2022										300
壁仞科技	壁砺™ 106C	2022										150
燧原科技	云燧 T20	2021							32	1600	300	300
燧原科技	云燧 T21	2021							32	1600	300	300
摩尔线程	MTT S3000	2022		4096	15.2				32	448		250
摩尔线程	MTT S4000	2023		8192	25	50	100	200	48	768		450

资料来源：各公司官网，海光信息招股说明书，寒武纪招股说明书，机器之心，中原证券研究所

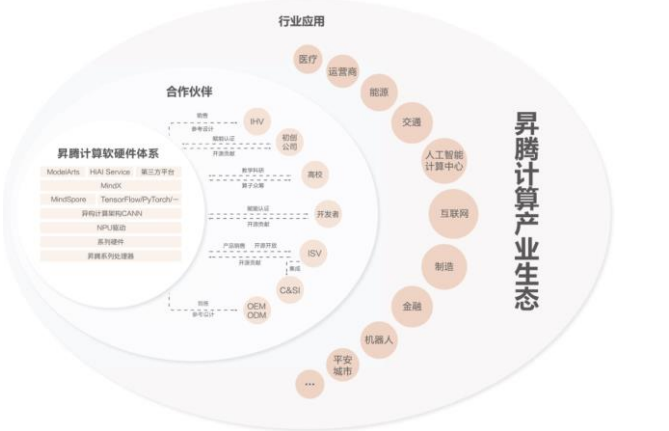
AI 算力芯片软件生态壁垒极高，国产领先厂商华为昇腾、寒武纪等未来有望在生态上取得突破。在软件生态方面，英伟达经过十几年的积累，其 CUDA 生态建立极高的竞争壁垒，国产厂商通过兼容 CUDA 及自建生态两条路径发展，国内领先厂商华为昇腾、寒武纪等未来有望在生态上取得突破。华为基于昇腾系列 AI 芯片，通过模组、板卡、小站、服务器、集群等丰富的产品形态，打造面向“端、边、云”的全场景 AI 基础设施方案。昇腾计算是基于硬件和基础软件构建的全栈 AI 计算基础设施、行业应用及服务，包括昇腾系列 AI 芯片、系列硬件、CANN（异构计算架构）、AI 计算框架、应用使能、开发工具链、管理运维工具、行业应用及服务等全产业链。昇腾计算已建立基于昇腾计算技术与产品、各种合作伙伴，为千行百业赋能的生态体系。

图 46：昇腾计算系统架构框图



资料来源：昇腾计算产业发展白皮书，中原证券

图 47：昇腾计算产业生态图



资料来源：昇腾计算产业发展白皮书，中原证券

3. DeepSeek 有望推动国产 AI 算力芯片加速发展

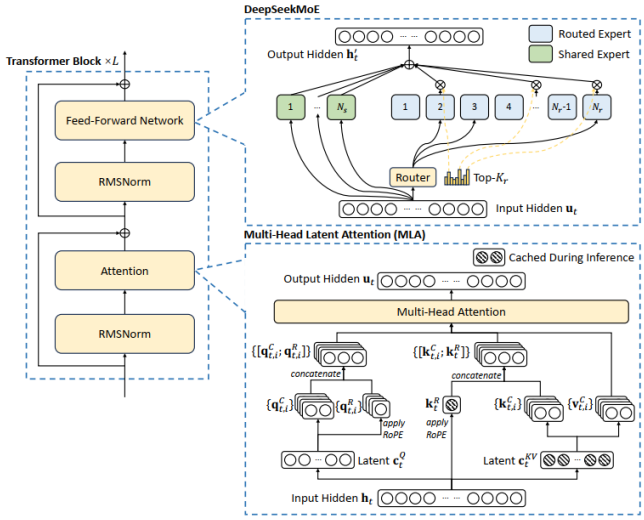
DeepSeek 通过技术创新实现大模型训练极高的性价比。2024 年 12 月 26 日，DeepSeek 正式发布全新系列模型 DeepSeek-V3，DeepSeek-V3 为自研 MoE 模型，总参数量为 671B，每个 token 激活 37B 参数，在 14.8T token 上进行了预训练。DeepSeek-V3 在性能上对标 OpenAI GPT-4o 模型，并在成本上优势巨大，实现极高的性价比。DeepSeek-V3 的技术创新主要体现在采用混合专家（MoE）架构，动态选择最合适的子模型来处理输入数据，以降低计算量；引入多头潜在注意力机制（MLA）降低内存占用和计算成本，同时保持高性能；采用 FP8 混合精度训练降低算力资源消耗，同时保持模型性能；采用多 Token 预测（MTP）方法提升模型训练和推理的效率。

DeepSeek MoE 架构通过动态组合多个专家模型来提升模型的性能和效率。DeepSeek 的 MoE 架构通过将传统 Transformer 中的前馈网络（FFN）层替换为 MoE 层，引入多个专家网络（Experts）和一个门控网络（Gating Network）。专家网络包括多个独立的专家模型，每个专家模型负责处理特定类型的数据。门控网络负责决定每个输入数据应该由哪些专家模型处理，并分配相应的权重；通过门控机制，模型能够动态选择最合适的专家来处理输入数据。DeepSeek MoE 架构采用稀疏激活策略，每次训练或推理时只激活部分专家，而不是整个模型；在 DeepSeek- V3 中，模型总参数为 6710 亿，但每次训练仅激活 370 亿参数，从而提高计算效率。传统的 Transformer 架构采用固定的编码器-解码器结构，所有输入数据通过相同的多层自注意力机制和前馈神经网络处理；模型的参数是静态的，无法根据输入数据的特性动态调整。

多头潜在注意力机制（MLA）的核心思想是对 KV 进行低秩压缩，以减少推理过程中的 KV 缓存，从而降低内存占用及计算成本。在传统的 Transformer 架构推理过程中，在进行生成式任务时，模型需要逐步生成序列，每次生成一个新 token 时，模型需要读入所有过去 Token 的上下文，重新计算之前所有 token 的键（Key）和值（Value）。KV 缓存通过存储这

些已计算的 Key 和 Value，避免重复计算，从而提高推理效率。MLA 的方法是将 KV 矩阵转换为低秩形式，将原矩阵表示为两个较小矩阵（相当于潜在向量）的乘积，在推理过程中，仅缓存潜在向量，而不缓存完整的 KV。这种低秩压缩技术显著减少了 KV 缓存的大小，同时保留了关键信息，从而降低内存占用及计算成本。

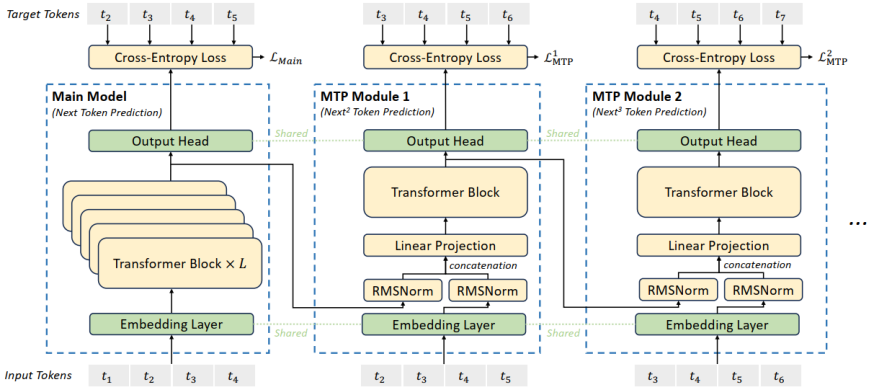
图 48：DeepSeek-V3 基本架构图



资料来源：DeepSeek-V3 Technical Report，中原证券研究所

多 token 预测（MTP）是一种创新的训练目标，通过同时预测多个未来 token 来提升模型的训练和推理效率。MTP 技术基于主模型（Main Model）和多个顺序模块（MTP Module），主模型负责基础的下一个 Token 预测，而 MTP 模块用于预测多个未来 Token。传统的模型通常一次只预测下一个 token，在生成文本时，模型按照顺序逐个生成下一个 Token，每生成一个 Token 都要进行一次完整的计算，依赖前一个生成的 Token 来生成下一个；而 MTP 能够同时预测多个连续的 Token，模型通过改造增加多个独立输出头，利用多 token 交叉熵损失进行训练，一次计算可以得到多个 Token 的预测结果，显著增加了训练信号的密度，提升模型的训练和推理效率，并且 MTP 生成的文本更加连贯自然，适合长文本生成任务。

图 49：DeepSeek-V3 MTP 应用示意图

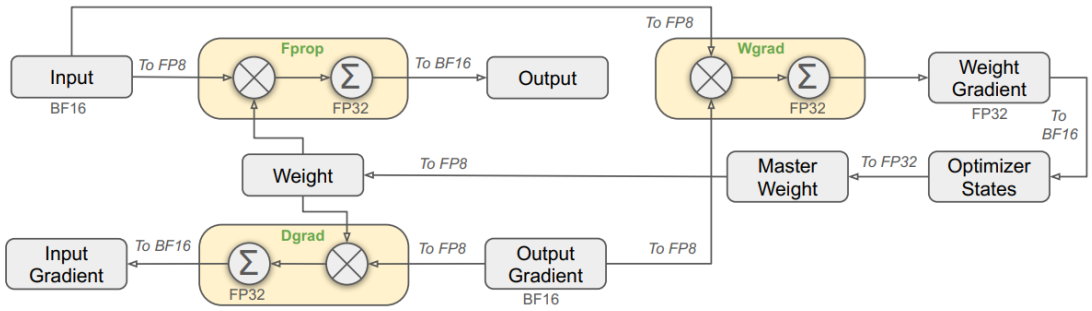


资料来源：DeepSeek-V3 Technical Report，中原证券研究所

DeepSeek 采用 FP8 混合精度训练技术在训练效率、内存占用和模型性能方面实现了显

著优化。传统大模型通常使用 FP32 或 FP16 进行训练，精度较高，但计算速度慢，内存占用较大。而 FP8 数据位宽是 8 位，与 FP16、FP32 相比，使用 FP8 进行计算的速度最快、内存占用最小。DeepSeek FP8 混合精度将 FP8 与 BF16、FP32 等结合，采用 FP8 进行大量核心计算操作，少数关键操作则使用 BF16 或 FP32，提高效率的同时确保数值稳定性，并显著减少了内存占用和计算开销。

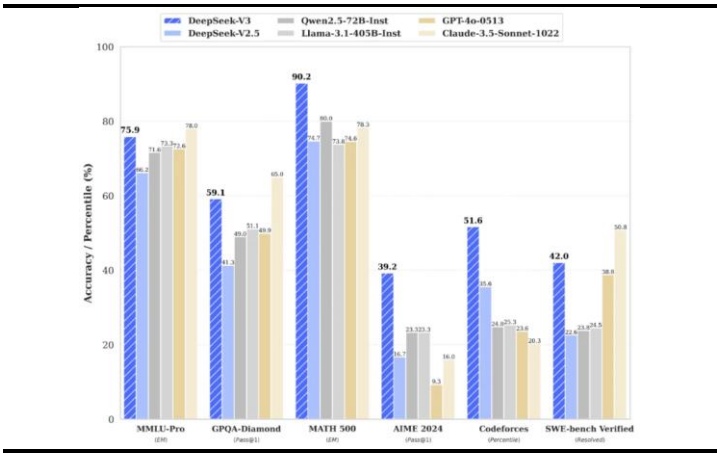
图 50：DeepSeek-V3 FP8 混合精度框架示意图



资料来源：DeepSeek-V3 Technical Report，中原证券研究所

DeepSeek-V3 性能对标 GPT-4o。DeepSeek-V3 多项评测成绩超越了 Qwen2.5-72B 和 Llama-3.1-405B 等其他开源模型，并在性能上和世界顶尖的闭源模型 GPT-4o 以及 Claude-3.5-Sonnet 不分伯仲。DeepSeek-V3 在知识类任务（MMLU, MMLU-Pro, GPQA, SimpleQA）上的水平相比前代 DeepSeek-V2.5 显著提升，接近当前表现最好的模型 Claude-3.5-Sonnet-1022；长文本测评方面，在 DROP、FRAMES 和 LongBench v2 上，DeepSeek-V3 平均表现超越其他模型；DeepSeek-V3 在算法类代码场景（Codeforces），远远领先于市面上已有的全部非 o1 类模型，并在工程类代码场景（SWE-Bench Verified）逼近 Claude-3.5-Sonnet-1022；在美国数学竞赛（AIME 2024, MATH）和全国高中数学联赛（CNMO 2024）上，DeepSeek-V3 大幅超过了所有开源闭源模型；DeepSeek-V3 与 Qwen2.5-72B 在教育类测评 C-Eval 和代词消歧等评测集上表现相近，但在事实知识 C-SimpleQA 上更为领先。

图 51：DeepSeek-V3 多项评测成绩对标 GPT-4o



资料来源：DeepSeek，中原证券研究所

图 52：DeepSeek-V3 多项评测成绩与其他大模型对比情况

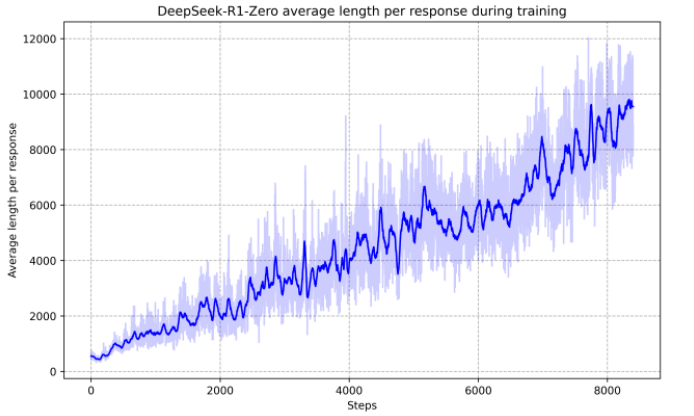
	测试集	DeepSeek-V3	Qwen2.5 72B-Inst.	Llama3.1 405B-Inst.	Claude-3.5-Sonnet-1022	GPT-4o 0513
	模型架构	MoE	Dense	Dense	-	-
	# 激活参数	37B	72B	405B	-	-
	# 总参数	671B	72B	405B	-	-
英文	MMLU (EM)	88.5	85.3	**88.6**	88.3	87.2
	MMLU-Redux (EM)	**89.1**	85.6	86.2	88.9	88
	MMLU-Pro (EM)	**75.9**	71.6	73.3	**78**	72.6
	DROP (3-shot F1)	**91.6**	76.7	88.7	88.3	83.7
	IF-Eval (Prompt Strict)	86.1	84.1	86	**86.5**	84.3
	GPQA-Diamond (Pass@1)	59.1	49	51.1	**65**	49.9
	SimpleQA (Correct)	24.9	9.1	17.1	28.4	**38.2**
	FRAMES (Acc.)	73.3	69.8	70	72.5	**80.5**
	LongBench v2 (Acc.)	**48.7**	39.4	36.1	41	48.1
代码	HumanEval-Mul (Pass@1)	**82.6**	77.3	77.2	81.7	80.5
	LiveCodeBench(Pass@1-COT)	**40.5**	31.1	28.4	36.3	33.4
	LiveCodeBench (Pass@1)	**37.6**	28.7	30.1	32.8	34.2
	Codeforces (Percentile)	**51.6**	24.8	25.3	20.3	23.6
	SWE Verified (Resolved)	42	23.8	24.5	**50.8**	38.8
	Aider-Edit (Acc.)	79.7	65.4	63.9	**84.2**	72.9
	Aider-Polyglot (Acc.)	**49.6**	7.6	5.8	45.3	16
数学	AIME 2024 (Pass@1)	**39.2**	23.3	23.3	16	9.3
	MATH-500 (EM)	**90.2**	80	73.8	78.3	74.6
	CNMO 2024 (Pass@1)	**43.2**	15.9	6.8	13.1	10.8
中文	CLUEWSC (EM)	90.9	**91.4**	84.7	85.4	87.9
	C-Eval (EM)	**86.5**	86.1	61.5	76.7	76
	C-SimpleQA (Correct)	**64.1**	48.4	50.4	51.3	59.3

资料来源：DeepSeek，中原证券研究所

DeepSeek-R1 通过冷启动与多阶段训练显著提升模型的推理能力，模型蒸馏技术有望推

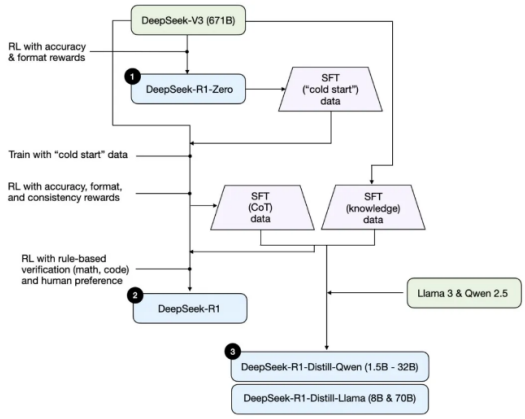
动 AI 应用加速落地。DeepSeek-R1-Zero 与 DeepSeek-R1 都是基于强化学习（RL）的推理模型，DeepSeek-R1-Zero 存在语言不一致等输出方面的问题，DeepSeek-R1 通过冷启动与多阶段训练，显著提升模型的推理能力，同时具有较好的实用性。DeepSeek-R1 采用模型蒸馏技术，将大模型（教师模型）的推理能力高效迁移到小模型（学生模型）中；模型蒸馏的核心思想是通过教师模型的输出指导学生模型的训练，使学生模型能够模仿教师模型的行为；通过蒸馏技术，小模型能够保留大模型的大部分性能，DeepSeek-R1 蒸馏后的小模型在多个基准测试中表现出色；DeepSeek-R1 的模型蒸馏技术显著提升小模型的推理能力，并降低部署成本，有望推动 AI 应用加速落地。

图 53：DeepSeek-R1-Zero 的思考时间持续提升以解决推理任务



资料来源：DeepSeek-R1 Technical Report，中原证券研究所

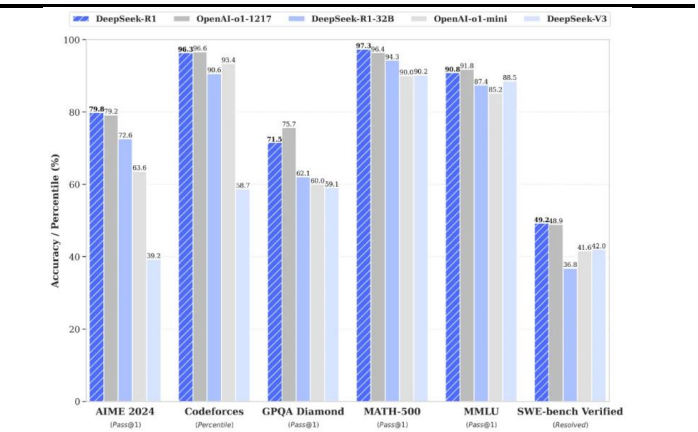
图 54：DeepSeek-R1-Zero、R1、蒸馏小模型的开发流程图



资料来源：机器之心，中原证券研究所

DeepSeek-R1 性能对标 OpenAI o1。DeepSeek-R1 极大提升了模型推理能力，在数学、代码、自然语言推理等任务上，性能比肩 OpenAI o1 正式版。DeepSeek 在开源 DeepSeek-R1-Zero 和 DeepSeek-R1 两个 660B 模型的同时，通过 DeepSeek-R1 的输出，蒸馏了 6 个小模型开源给社区，其中 32B 和 70B 模型在多项能力上实现了对标 OpenAI o1-mini 的效果。

图 55：DeepSeek-R1 多项评测成绩对标 OpenAI o1



资料来源：DeepSeek，中原证券研究所

图 56：DeepSeek-R1 蒸馏 32B 和 70B 模型多项评测成绩对标 OpenAI o1-mini

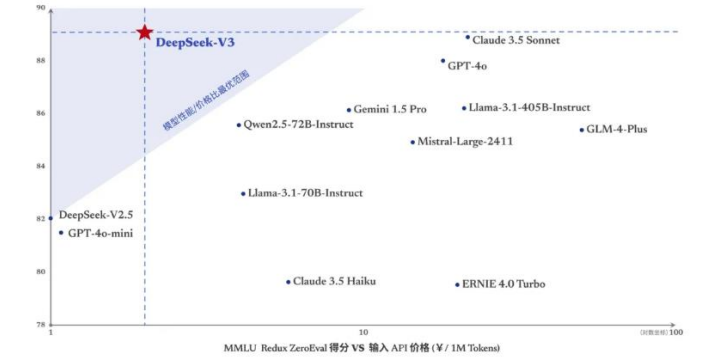
	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759.0
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717.0
o1-mini	63.6	80.0	90.0	60.0	53.8	**1820.0**
QwQ-32B	44.0	60.0	90.6	54.5	41.9	1316.0
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954.0
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189.0
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481.0
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691.0
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205.0
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633.0

资料来源：DeepSeek，中原证券研究所

DeepSeek 实现大模型训练与推理成本优势巨大，助力 AI 应用大规模落地。DeepSeek-V3 的训练成本具有极大的经济性，根据 DeepSeek-R1 Technical Report 的数据，在预训练阶

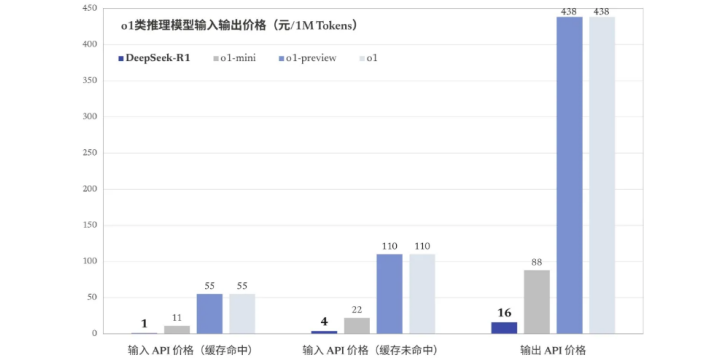
段，每处理 1 万亿 tokens，训练 DeepSeek-V3 仅需 18 万 H800 GPU 小时，即在 2048 块 H800 GPU 的集群上需要 3.7 天；因此，DeepSeek-V3 的预训练阶段在不到两个月内完成，耗时 266.4 万（2664K）GPU 小时；加上上下文长度扩展所需的 11.9 万 GPU 小时和后训练所需的 5 千 GPU 小时，DeepSeek-V3 的完整训练仅需 278.8 万 GPU 小时；假设 H800 GPU 的租赁价格为每小时 2 美元，DeepSeek-V3 的总训练成本仅为 557.6 万美元。2025 年 1 月 20 日 DeepSeek-R1 正式发布，其 API 定价为每百万输入 tokens 1 元（缓存命中）/ 4 元（缓存未命中），每百万输出 tokens 16 元；OpenAI o1 定价为每百万输入 tokens 55 元（缓存命中）/110 元（缓存未命中），每百万输出 tokens 438 元；DeepSeek-R1 API 调用成本不到 OpenAI o1 的 5%。DeepSeek-V3 性能对标 GPT-4o，DeepSeek-R1 性能对标 OpenAI o1，并且 DeepSeek 模型成本优势巨大，有望推动 AI 应用大规模落地。

图 57：DeepSeek-V3 模型性价比处于最优范围



资料来源：DeepSeek，中原证券研究所

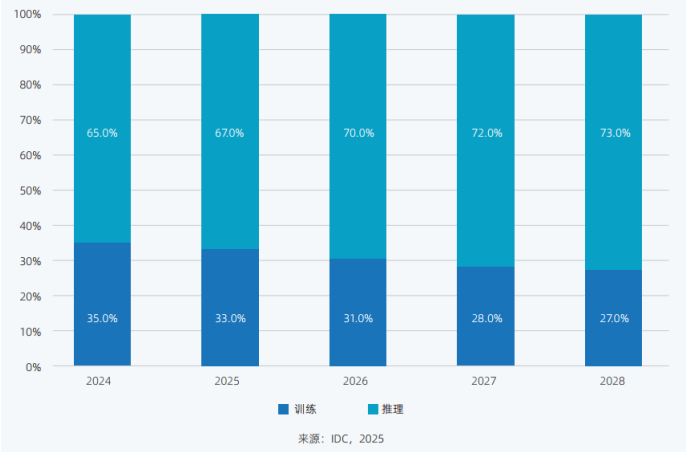
图 58：DeepSeek-R1 与 OpenAI o1 类推理模型 API 定价对比情况（2025 年 1 月 20 日）



资料来源：DeepSeek，中原证券研究所

DeepSeek 有望推动推理需求加速释放，国产 AI 算力芯片或持续提升市场份额。随着大模型的成熟及 AI 应用的不断拓展，推理场景需求日益增加，推理服务器的占比将显著提高；IDC 预计 2028 年中国 AI 服务器用于推理工作负载占比将达到 73%。根据的 IDC 数据，2024 上半年，中国加速芯片的市场规模达超过 90 万张，国产 AI 芯片出货量已接近 20 万张，约占整个市场份额的 20%；用于推理的 AI 芯片占据 61%的市场份额。DeepSeek-R1 通过技术创新实现模型推理极高性价比，蒸馏技术使小模型也具有强大的推理能力及低成本，将助力 AI 应用大规模落地，有望推动推理需求加速释放。由于推理服务器占比远高于训练服务器，在 AI 算力芯片进口受限的背景下，用于推理的 AI 算力芯片国产替代空间更为广阔，国产 AI 算力芯片有望持续提升市场份额。

图 59：2024-2028 年中国 AI 服务器工作负载预测情况



资料来源：IDC，2025 中国人工智能计算力发展评估报告，中原证券研究所

国产算力生态链全面适配 DeepSeek，国产 AI 算力芯片厂商有望加速发展。DeepSeek 大模型得到全球众多科技厂商的认可，纷纷对 DeepSeek 模型进行支持，国内 AI 算力芯片厂商、CPU 厂商、操作系统厂商、AI 服务器及一体机厂商、云计算及 IDC 厂商等国产算力生态链全面适配 DeepSeek，有望加速 AI 应用落地。华为昇腾、沐曦、天数智芯、摩尔线程、海光信息、壁仞科技、寒武纪、云天励飞、燧原科技、昆仑芯等国产 AI 算力芯片厂商已完成适配 DeepSeek，DeepSeek 通过技术创新提升 AI 算力芯片的效率，进而加快国产 AI 算力芯片自主可控的进程，国产 AI 算力芯片厂商有望加速发展。

表 8：官宣支持 DeepSeek 模型的国产 AI 芯片企业动态

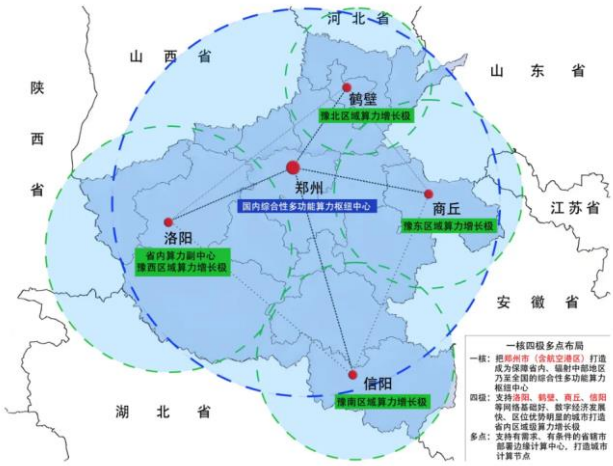
公司	日期	支持情况
华为	2 月 1 日	首发！硅基流动 x 华为云联合推出基于昇腾云的 DeepSeek R1&V3 推理服务！
沐曦	2 月 1 日	Gitee AI 联合沐曦首发全套 DeepSeek R1 千问蒸馏模型，全免费体验！
天数智芯	2 月 4 日	一天适配！天数智芯联合 GiteeAI 正式上线 DeepSeek
摩尔线程	2 月 4 日	致敬 DeepSeek：以国产 GPU 为基，燎原中国 AI 生态之火
海光信息	2 月 4 日	DeepSeek V3 和 R1 模型完成海光 DCU 适配并正式上线
壁仞科技	2 月 5 日	DeepSeek R1 在壁仞国产 AI 算力平台发布，全系列模型一站式赋能开发者创新
太初元碁	2 月 5 日	基于太初 T100 加速卡 2 小时适配 DeepSeek-R1 系列模型
云天励飞	2 月 5 日	DeepEdge10 已完成 DeepSeek R1 系列模型适配
燧原科技	2 月 6 日	燧原科技实现全国各地智算中心 DeepSeek 的全量推理服务部署
昆仑芯	2 月 6 日	国产 AI 卡 Deepseek 训练推理全版本适配、性能卓越
灵汐科技	2 月 6 日	灵汐芯片快速实现 DeepSeek 适配，助力国产大模型与类脑智能硬件融合
鲲云科技	2 月 6 日	鲲云科技 CAISA 430 适配 DeepSeek R1 推理，开启高效 AI 应用新时代
希姆计算	2 月 6 日	希姆计算开源算力全面适配 DeepSeek-R1 开源模型
寒武纪	2 月 6 日	南京智算中心与寒武纪、苏宁科技合作，成功上线全国产算力版 DeepSeek
算能	2 月 7 日	最佳国产边缘部署方案！DeepSeek-R1 蒸馏模型已适配 SE7，代码全开源！
清微智能	2 月 7 日	清微智能可重构算力芯片全面适配 DeepSeek 模型推理和训练
芯动力	2 月 7 日	芯动力神速适配 DeepSeek-R1 大模型，AI 芯片设计迈入“快车道”！
墨芯	2 月 7 日	墨芯 S40 计算卡完成 DeepSeek 大模型部署，支持单卡推理大模型
后摩智能	2 月 7 日	开源破局 x 低功耗守护：Deepseek 与存算一体如何演绎 AI 界的"哪吒闹海"？
瀚博	2 月 8 日	瀚博完成 DeepSeek 全版本训推适配，单机支持 V3 与 R1 671B 满血版部署
爱芯元智	2 月 8 日	爱芯分享|基于 AX650N&AX630C 部署 DeepSeek R1
芯瞳	2 月 9 日	芯瞳 GPU 完成与 DeepSeek 的适配，向中国 AI 开发者致敬
进迭时空	2 月 10 日	进迭时空 Bianbu Cloud 成功运行 DeepSeek 本地大模型
江原科技	2 月 11 日	江原科技实现全国产 AI 推理芯片单卡支持 DeepSeek-R1-70B 部署
奕斯伟	2 月 14 日	奕斯伟计算|技术新突破！RISC-V AI SoC 成功适配 DeepSeek 模型计算

资料来源：芯东西，中原证券研究所

4. 河南省着力布局 AI 算力芯片，产业链初具雏形

河南省以“一核四极多点”为核心框架进行算力产业布局，打造全国重要算力高地。河南省将算力作为支撑数字河南建设的重要底座和驱动数字化转型的新引擎，致力于打造面向中部、辐射全国的算力调度核心枢纽和全国重要的算力高地。河南省的算力产业布局以“一核四极多点”为核心框架，以郑州市（含航空港区）为核心，依托其网络枢纽地位和算力资源，构建国家超算互联网核心节点和智算中心集群，打造综合性多功能算力枢纽中心；支持洛阳、鹤壁、商丘、信阳等城市作为区域增长极，利用当地算力资源，面向周边区域提供算力服务；鼓励有条件的地方部署边缘计算中心，打造城市计算节点，满足本地业务需求。到 2026 年，河南省计划形成布局合理、绿色低碳、高效集约、安全可靠的算力基础设施格局，全省算力基础设施标准机架数达到 35 万架，平均利用率达到 70%以上，算力规模超过 120EFlops，其中智算、超算等高性能算力占比超过 90%。

图 60：河南省 “一核四极多点”算力产业布局示意图



资料来源：河南省发改委，中原证券研究所

河南空港智算中心将打造成为“算力+产业”标杆。河南空港智算中心项目为中部地区规模最大的智算中心，开建于 2024 年 6 月，按照 A 级数据中心标准建设 15 个模块化机房，主要满足大模型研发企业的高端训练算力需求，仅用百天即完成了首期 2000P（1P 约等于每秒 1000 万亿次浮点运算能力）算力部署。2025 年一季度，计划该项目算力投产规模可达 10000P，项目一期全部建成后，将达到 30000P 算力规模，为郑州航空港科技创新和产业升级提供强有力支撑。河南空港智算中心作为中部首个同时部署全量级 DeepSeek-V3/R1 及多模态 DeepSeek-Janus-Pro 模型的机构，基于 DeepSeek-R1 打造的首个企业级 AI 办公智能体应用已正式投入使用，DeepSeek-V3/R1 的部署将极大地推动 AI 大模型在医疗、教育、科研、工业、无人驾驶、智慧城市、交通物流、游戏、视频等领域的广泛应用，为各行各业的发展注入强大动力。河南空港智算中心所运营的产业园区重点聚焦数字经济、新一代信息技术及智能制造高端服务产业，产业园区以河南空港智算中心为基座，构建“1 个智算中枢+N 个垂直场景”产业架构，通过“链主牵引+生态协同”模式，目前已吸引了新华三、科大讯飞、腾讯云等 40 余家创新企业入驻。

图 61：河南空港智算中心示意图

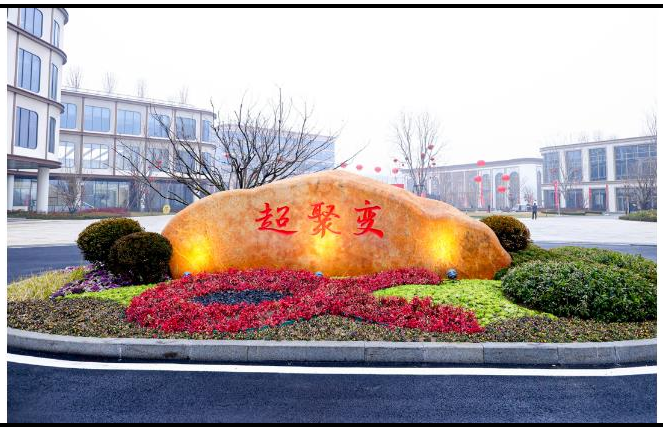
资料来源：河南郑州航空港发布，中原证券研究所

图 62：算力将赋能千行百业

资料来源：河南省发改委，中原证券研究所

依托省内先进计算企业，构建算力产业生态。河南省依托超聚变研发中心及总部基地、紫光智慧终端产业园等重大项目，积极引进芯片等上游企业，吸引集聚服务器操作系统、数据库、中间件开发骨干企业，打造先进计算产业园、鲲鹏软件小镇等园区，构建具有国际竞争力的先进计算产业集群。超聚变在中国服务器市场稳居第二，AI 服务器市场位居第一，2024 年营收达 400 亿，海外市场三年复合增长率超过 50%，合作伙伴数量已达 22000 家。超聚变研发中心及总部基地是河南省算力产业的重要项目，于 2025 年 3 月 1 日正式启用，该项目将助力超聚变在全球范围内开展日常运营及产品研发。超聚变计划通过总部基地构建本土产业链生态，推动河南制造走向全球，参与全球算力产业分工。

图 63：超聚变研发中心及总部基地



资料来源：超聚变，中原证券研究所

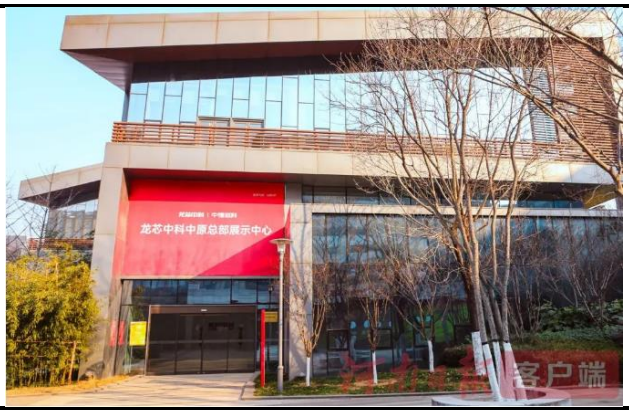
图 64：超聚变稳居中国服务器市场第二

资料来源：超聚变，中原证券研究所

河南省着力布局 AI 算力芯片，产业链初具雏形。2022 年 8 月，龙芯中科技术股份有限公司与河南省政府签署战略合作协议，将在河南建设龙芯生态，并在鹤壁形成产业积聚。2023 年龙芯中科芯片封装基地一期在鹤壁正式投产，具备龙芯一号芯片封装、测试和出货的能力，整个项目建成达产后，可实现年封装测试芯片 3000 万片。随着龙芯中科鹤壁产业基地产能逐步释放，已有众多上下游企业在当地形成积聚，目前已经引进了云涌科技、力积存储等 12 家硬件生产企业，麒麟、统信等 10 家软件企业，为河南人工智能算力产业链的发展提供硬件和软件支持。2023 年 11 月龙芯中科中原总部基地在郑州航空港经济综合实验区揭牌，龙芯中科中原总部基地将建设研发创新中心、生态适配中心、信创展示中心等，也将为河南人工智能算力产业发展提供技术研发和生态适配支持。沐曦致力于为异构计算提供全栈 GPU 芯片及

解决方案，可广泛应用于智算、智慧城市、云计算、自动驾驶、数字孪生、元宇宙等前沿领域，为数字经济发展提供强大的算力支撑。2025 年 2 月 5 日沐曦与联想集团联合发布首个国产 DeepSeek 一体机解决方案，截止 2025 年 3 月 7 日，该解决方案累计发货量已突破千台，配备沐曦国产 GPU 卡近万张，覆盖医疗、教育、制造等十余个核心行业，标志着国产 AI 产业落地的重要里程碑。河南投资集团通过算力产业基金投资沐曦集成，推动沐曦集成在河南落地，助力算力产业生态的构建。河南省通过引进、投资、培育本土企业等方式布局 AI 算力芯片，产业链初具雏形。

图 65：龙芯中科中原总部



资料来源：河南郑州航空港发布，中原证券研究所

图 66：联想沐曦 DeepSeek 一体机

资料来源：新浪，中原证券研究所

河南省政策大力扶持 AI 算力芯片产业。2024 年 11 月 7 日，《河南省算力基础设施发展规划（2024—2026 年）》正式发布，规划提出要培育人工智能产业，突破发展人工智能芯片，吸引集聚一批人工智能相关软件及服务、芯片研发制造等企业；推动技术创新，强化算力领域学术界与产业界的交流合作，聚焦大规模数据处理、内存计算、异构计算、存算一体、算网融合等关键共性技术开展研发攻关，支持企业建设算力领域研发创新平台，引导企业加大人工智能服务器、计算芯片、人工智能软件等研发投入，布局发展国产高性能计算软件系统、国产数据库，提升关键配套能力；加大资金支持力度，强化财政资金引导作用，统筹各类相关专项资金重点支持算力基础设施建设、算力产业发展以及算力生态搭建，鼓励银行将算力列为科技信贷业务重点支持领域，支持符合条件的企业通过发行绿色债券或上市实现融资；增强算力设施可靠性，鼓励智算、超算中心采用昇腾、海光等自主可控技术路线。

表 9：2021-2024 年河南省人工智能产业部分重要产业政策情况

时间（年）	发布单位	政策名称	政策主要内容
2021	省政府办公厅	《河南省推进新型基础设施建设行动计划（2021—2023年）》	该计划提出要建设全栈国产化、自主可控智能计算中心，打造一批公共数据资源库、标注数据库、训练数据库、开源训练数据集等基础平台，完善智能算力基础设施；建设全省统一的智能网联汽车云控平台，开展中原科技城自动驾驶公交线路示范应用；支持郑州市创建国家新一代人工智能创新发展试验区。
2021	省政府办公厅	《河南省“十四五”战略性新兴产业和未来产业发展规划》	该规划提出加强人工智能领域基础理论研究与关键共性技术攻关，重点突破图像识别感知、数字图像处理、语音识别、智能判断决策等核心应用技术，引进一批人工智能龙头企业，加快培育壮大本地企业，做强智能网联汽车、智能机器人、智能无人机、智能计算设备、智能家居产品等优势智能产品；深化人工智能技术在智能制造、现代农业、智慧城市、智慧文旅、智慧医疗等领域的创新应用，创建国家新一代人工智能创新发展试验区。
2022	省政府办公厅	《河南省“十四五”战略性新兴产业和未来产业发展规划》	该规划提出新一代信息技术产业聚焦“补芯、引屏、固网、强端、育器”，强化信息制造、信息基础设施和信息安全等重点领域创新，推动大数据、人工智能、区块链等技术和实体经济深度融合，构建万物互联、融合创新、智能协同、绿色安全的产业发展生态。到 2025 年，新一代信息技术产业营业收入超过 1 万亿元。。
2022	郑州人民政府	《郑州国家新一代人工智能创新发展试验区建设实施方案》	该方案提出要培育人工智能创新企业，培育 30 家人工智能创新标杆企业，形成 5~10 家在国内人工智能领域具有影响力的一流创新主体；设立人工智能创新发展专项资金，重点支持人工智能产业的基础研究、关键共性技术攻关、场景应用示范等；统筹利用省、市高端人才计划，引进培育 20 个人工智能高层次领军人才团队。
2024	省政府办公厅	《河南省算力基础设施发展规划（2024—2026 年）》	该规划提出要培育人工智能产业，突破发展人工智能芯片，吸引集聚一批人工智能相关软件及服务、芯片研发制造等企业；推动技术创新，强化算力领域学术界与产业界的交流合作，引导企业加大人工智能服务器、计算芯片、人工智能软件等研发投入，布局发展国产高性能计算软件系统、国产数据库，提升关键配套能力；加大资金支持力度，强化财政资金引导作用；增强算力设施可靠性，鼓励智算、超算中心采用昇腾、海光等自主可控技术路线。
2024	省政府办公厅	《河南省推动“人工智能+”行动计划（2024—2026年）》	该规划提出到 2026 年年底，力争 2—3 个行业人工智能应用走在全国前列，建设一批高质量行业数据集，形成 2—3 个先进可用的基础大模型、20 个以上垂直领域行业模型和一批面向细分场景的应用模型、100 个左右示范引领典型案例，涌现一批制度创新典型做法和服务行业应用的标准规范；探索人工智能在能源、金融、人力资源、消费等行业多元化应用，形成人工智能行业应用新生态。

资料来源：省政府办公厅，郑州人民政府，中原证券研究所

5. 河南省 AI 算力芯片产业相关企业

5.1. 龙芯中科

龙心中科为国产处理器领先企业，建立自主可控生态体系。公司成立于 2008 年，坚持自主研发，推出自主指令系统龙架构，持续研发及优化多个自主软/硬 IP 核，不依赖国外技术授权（包括指令系统、IP 核等），不依赖境外供应链，从基于自主 IP 的芯片研发、基于自主工艺的芯片生产、基于自主指令系统的软件生态三个环节提高自主可控度，保障供应链安全的同时基于自主技术构建自主体系。公司是国内 CPU 企业中极个别可以进行指令系统架构及 CPU IP 核授权的企业，是极个别在股权结构方面保持开放、未被整机厂商控制的企业。

公司掌握核心技术并持续建设产业生态，构筑核心竞争力。龙芯中科是国内唯一坚持基

于自主指令系统构建独立于 Wintel 体系和 AA 体系的开放性信息技术体系和产业生态的 CPU 企业。公司坚持自主研发核心 IP，形成了包括系列化 CPU IP 核、GPU IP 核、内存控制器及 PHY、高速总线控制器及 PHY 等上百种 IP 核。公司推出了自主指令系统 LA，并基于 LA 迁移或研发了操作系统的核心模块，包括内核、三大编译器（GCC、LLVM、GoLang）、三大虚拟机（Java、Java Script、.NET）、浏览器、媒体播放器、KVM 虚拟机等，形成了面向服务器、面向桌面和面向工控类应用的基础版操作系统。公司通过设计优化和先进工艺提升性能，摆脱对最先进工艺的依赖。公司自主研发了包括处理器核心在内的上百种核心模块，产品竞争力不断提升与市场应用持续辐射产业链，目前与公司开展合作的厂商达到数千家，下游开发人员达到数十万人，基于龙芯处理器的自主信息产业生态体系正在逐步形成。

公司 CPU 产品性能突出，覆盖信息化、工控市场。公司处理器及配套芯片产品包括龙芯 1 号、龙芯 2 号、龙芯 3 号三大系列处理器芯片及桥片等配套芯片，面向工控等领域的 2K0300 嵌入式 SoC 研制成功，面向服务器领域的 3C6000 处理器芯片样片研制成功；公司 3A6000 在桌面领域性能达到市场主流桌面 CPU 水平，3C6000 在服务器领域性能将达到市场主流服务器 CPU 水平，并具有性价比优势。公司基于开放的龙芯生态体系，与板卡、整机厂商及基础软件、应用解决方案开发商建立紧密的合作关系，为下游企业提供基于龙芯处理器的各类开发板及软硬件模块，并提供完善的技术支持与服务。公司持续强化 PC 和服务器主板 ODM 能力，与 CPU、操作系统形成“三位一体”能力。2024 年上半年，在信息化应用领域公司联合优质 ODM 企业推出丰富多样的 3A6000 产品解决方案，包括台式机、一体机、笔记本、NUC 等；服务器 CPU 方面，支持下游厂家完成龙芯 3C5000/3D5000 双路与四路服务器研制，进入市场推广阶段，基于龙芯 CPU 的服务器入围中国移动等运营商服务器集采标包。2025 年 2 月，搭载龙芯 3 号 CPU 的设备成功启动运行 DeepSeek R1 7B 模型，实现本地化部署，性能卓越，成本优异。

公司掌握 GPU 设计技术，已布局 AI 加速芯片。公司掌握图形处理器设计技术，可实现传统图形管线与大规模并行计算相结合的统一渲染架构，支持图形处理和通用计算加速。2024 年上半年，公司在支持图形渲染与通用计算的龙芯第二代图形处理器核上持续投入，在 2K3000 平台中完成 LG200 GPU 核的硅前验证工作，并交付流片。公司首款独立显卡/AI 加速卡芯片 9A1000 的研制工作全面展开，其图形处理器核在原有基础上进行功能、性能扩展，同时通过设计优化提高单位面积性能。

2024 年，由于传统优势工控市场停滞影响仍存在，工控类芯片营收大幅下降；电子政务市场开始回暖，信息化类芯片收入大幅增加；芯片类产品营收同比有较大幅度增长的同时，公司主动减少解决方案类业务，解决方案类业务营收同比有较大幅度下降。2024 年公司实现营业收入 5.07 亿元，同比增长 0.24%；实现归母净利润-6.24 亿元。

5.2. 沐曦

沐曦致力于为异构计算提供全栈 GPU 芯片及解决方案。沐曦集成电路（上海）股份有限公司成立于 2020 年 9 月，拥有技术完备、设计和产业化经验丰富的团队，核心成员平均拥有

近 20 年高性能 GPU 产品端到端研发经验，曾主导过十多款世界主流高性能 GPU 产品研发及量产，包括 GPU 架构定义、GPU IP 设计、GPU SoC 设计及 GPU 系统解决方案的量产交付全流程。公司致力于为异构计算提供全栈 GPU 芯片及解决方案，可广泛应用于智算、智慧城市、云计算、自动驾驶、数字孪生、元宇宙等前沿领域，为数字经济发展提供强大的算力支撑。

公司 GPU 产品拥有自主知识产权，覆盖智算推理、通用计算、图形渲染市场。沐曦打造全栈 GPU 芯片产品，推出曦思 N 系列 GPU 产品用于智算推理，曦云 C 系列 GPU 产品用于通用计算，以及曦彩 G 系列 GPU 产品用于图形渲染，满足“高能效”和“高通用性”的算力需求。沐曦产品均采用完全自主研发的 GPU IP，拥有完全自主知识产权的指令集和架构，配以兼容主流 GPU 生态的完整软件栈（MXMACA），具备高能效和高通用性的天然优势，能够为客户构建软硬件一体的全面生态解决方案，是“双碳”背景下推动数字经济建设和产业数字化、智能化转型升级的算力基石。

沐曦快速适配 DeepSeek 大模型，DeepSeek 一体机需求旺盛。2025 年 2 月 5 日，联想集团与沐曦联合发布首个国产 DeepSeek 一体机解决方案，该解决方案以“联想服务器/工作站+沐曦训推一体国产 GPU+自主算法”为核心架构为优势，覆盖主流用户场景，其搭载的异构计算架构可支持需要大量数据处理的场景，全面覆盖模型推理、模型训练、知识库管理和智能体开发四大开发场景，以及智慧办公、代码开发、客户服务、公文写作、视频生成及智能体实训教育等六大用户应用场景。自 DeepSeek 一体机面市以来，各行业本地部署大模型的需求持续攀升，截至 2025 年 3 月 7 日，该解决方案累计发货量已突破千台，配备沐曦国产 GPU 卡近万张，覆盖医疗、教育、制造等十余个核心行业，标志着国产 AI 产业落地的重要里程碑。

行业投资评级

强于大市：未来 6 个月内行业指数相对沪深 300 涨幅 10%以上；
同步大市：未来 6 个月内行业指数相对沪深 300 涨幅－10%至 10%之间；
弱于大市：未来 6 个月内行业指数相对沪深 300 跌幅 10%以上。

公司投资评级

买入：　未来 6 个月内公司相对沪深 300 涨幅 15%以上；
增持：　未来 6 个月内公司相对沪深 300 涨幅 5%至 15%；
谨慎增持：未来 6 个月内公司相对沪深 300 涨幅－10%至 5%；
减持：　未来 6 个月内公司相对沪深 300 涨幅－15%至－10%；
卖出：　未来 6 个月内公司相对沪深 300 跌幅 15%以上。

证券分析师承诺

本报告署名分析师具有中国证券业协会授予的证券分析师执业资格，本人任职符合监管机构相关合规要求。本人基于认真审慎的职业态度、专业严谨的研究方法与分析逻辑，独立、客观的制作本报告。本报告准确的反映了本人的研究观点，本人对报告内容和观点负责，保证报告信息来源合法合规。

重要声明

中原证券股份有限公司具备证券投资咨询业务资格。本报告由中原证券股份有限公司（以下简称“本公司”）制作并仅向本公司客户发布，本公司不会因任何机构或个人接收到本报告而视其为本公司的当然客户。

本报告中的信息均来源于已公开的资料，本公司对这些信息的准确性及完整性不作任何保证，也不保证所含的信息不会发生任何变更。本报告中的推测、预测、评估、建议均为报告发布日的判断，本报告中的证券或投资标的价格、价值及投资带来的收益可能会波动，过往的业绩表现也不应当作为未来证券或投资标的表现的依据和担保。报告中的信息或所表达的意见并不构成所述证券买卖的出价或征价。本报告所含观点和建议并未考虑投资者的具体投资目标、财务状况以及特殊需求，任何时候不应视为对特定投资者关于特定证券或投资标的的推荐。

本报告具有专业性，仅供专业投资者和合格投资者参考。根据《证券期货投资者适当性管理办法》相关规定，本报告作为资讯类服务属于低风险（R1）等级，普通投资者应在投资顾问指导下谨慎使用。

若本公司客户（以下简称“该客户”）向第三方发送本报告，则由该客户独自为其发送行为负责，提醒通过该种途径获得本报告的投资者注意，本公司不对通过该种途径获得本报告所引起的任何损失承担任何责任。

特别声明

在合法合规的前提下，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头寸并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问等各种服务。本公司资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告意见或者建议不一致的投资决策。投资者应当考虑到潜在的利益冲突，勿将本报告作为投资或者其他决定的唯一信赖依据。