国投证券 SDIC SECURITIES

2024 年 01 月 17 日
电子

行业专题

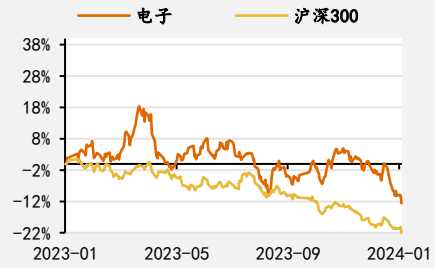
证券研究报告

AI 浪潮势不可挡，昇腾发力铸造国产算力基石

投资评级 领先大市-A
维持评级

首选股票	目标价（元）	评级

行业表现



资料：Wind 资讯

升幅%	1M	3M	12M
相对收益	-7.9	-0.3	9.4
绝对收益	-11.2	-11.6	-12.6

相关报告

CES 展会新品纷呈，AI 与显示交互成为焦点	2024-01-14
高端 系列--光刻胶：半导体制造核心材料，突围在即	2024-01-10
高通推出第二代骁龙 XR2+平台，原生鸿蒙生态有望迎来快速发展	2024-01-07
、小米新车相继发布，全球首款商务 AI PC 亮相	2024-01-01
苹果 MR 发售在即，美光业绩超预期	2023-12-24

从“性能指标”到“性能密度指标”，海外高端芯片进口受限范围扩大：

2023 年 10 月 17 日，美国商务部出台了出口 清单的 ECNN 3A090 和 4A090 要求，以进一步限制高性能 AI 芯片的出口，同时将 13 家中国公司列入实体清单。在新规发布之前，超过旧规性能指标限制的芯片仅为英伟达 A100，但当加入性能密度指标后，新规不仅限制了厂商出略低于性能标准的芯片以规避限制情况，同时针对数据中心芯片与非数据中心芯片进行了不同的限制约定，使更多的英伟达芯片受到禁令限制。修改后的出国 设计产品包括但不限于：英伟达 A100、A800、H100、H800、L40、L40S 以及 RTX 4090 产品。实际上，任何集成了一个或多个及以上的芯片的系统，包括但不限于英伟达 DGX、HGX 系统，都在新规涵盖范围之内。

全球科技巨头纷纷布局算力芯片，AI 浪潮势不可挡：

AMD 推出的“MI 系列+Infinity Fabric+ROCm 平台”，性能强劲，成为英伟达全球范围内最强劲的对手。特斯拉自研 Dojo 超算服务器助力自动驾驶，在芯片间互连技术上独具特色，具备强大的扩展性。Intel 推出 GPU 系列芯片和 oneAPI 开发平台，不断完善其 AI 数据中心布局。Google 推出 Cloud TPU 解决方案，在机器学习领域。Meta 布局自研 AI 生态，2020 年正式推出第一代 MTIA 方案，侧重于处理低/中复杂度模型。英伟达在 GPU 领域深耕数十年，根据 Jon Peddie Research（JPR）报告显示，2023 年 Q1，英伟达 GPU 市场份额达到 84%，Q2 达到 80%，但全球各大科技巨头结合其自身优势纷纷布局算力芯片。

完整的芯片生态“软件+硬件”缺一不可， 推出全系列昇腾解决方案：

一个完整的芯片生态中仅有硬件芯片是不够的，软件系统开发的便捷性以及现有系统的迁移难度同样至关重要。现阶段英伟达通过其“芯片+NVLINK+CUDA”构建产品高护城河，在 GPU 市场长期占据全球市场领导地位。 推出全系列昇腾解决方案，包括自研的昇腾芯片+高速芯片间接口+AI 计算架构，从软硬件上和主流厂商对标，昇腾芯片基于自研的达芬奇架构，功耗低，产品性能优异，从官方披露的数据来看，2019 年推出的昇腾 910，在半精度 FP16 运算速度可达

320TFLOPS，整型 INT8 算力可达 640TOPS，英伟达 A100 的 FP16 运算性能为 312TFLOPS。

相关公司：兴森科技、新益昌、天承科技、德邦科技、华海诚科、英维克、飞荣达、思泉新材、恒铭达、华丰科技、沪电股份、世运电路、胜宏科技、方正科技。

风险提示：宏观因素波动影响下游需求；市场开拓不及预期；行业竞争加剧；

内容目录

图表目录

1. 美国加强限制规则，海外高性能芯片进口受限

1.1. 从“性能指标”到“性能密度指标”，英伟达高端芯片进口受限范围扩大

2023 年 10 月 17 日，美国商务部出台了出口　　清单的 ECNN 3A090 和 4A090 要求，以进一步限制高性能 AI 芯片的出口，同时将 13 家中国公司列入实体清单。修改后的出国　　设计产品包括但不限于：英伟达 A100、A800、H100、H800、L40、L40S 以及 RTX 4090 产品。实际上，任何集成了一个或多个及以上的芯片的系统，包括但不限于英伟达 DGX、HGX 系统，都在新规涵盖范围之内。

此前，2022 年 8 月 26 日，美国政府要求英伟达停止向中国（包括中国香港）出口两款用于人工智能发展的高端计算芯片，涉及英伟达 A100 和 H100 两款芯片，以及未来推出峰值性能等同或超过 A100 的其他芯片。同时，英伟达应用这些高性能芯片的系统级产品也均在新的　　范围内。2022 年 9 月 1 日，英伟达发布声明称美国政府允许英伟达在 2023 年 9 月 1 日前，通过公司的香港工厂履行 A100 和 H100 的订单和物流运输，但售卖给中国的终端客户仍需要受美国政府批准。

表1：限制法令变化

内容	2022 年 10 月 7 日	2023 年 10 月 17 日新规
性能指标	限制每次运算位长乘以 TOPS 为单位的处理性能不能大于 4800	增加性能密度指标，同时限制总处理性能与性能密度
限制主体	拒绝英伟达向中国出口、转让的申请	扩大限制范围，不仅包括大陆与澳门特别行政区，还包括母公司设置在中国大陆以及澳门特别行政区的实体

资料　　：《对向中国出口的先进计算和半导体制造物项实施新的出口　　》(2022)，《先进计算芯片更新规则》及 《半导体制造物项更新规则》(2023)，国投证券研究中心

限制强度加大，新增多款芯片受到新规限制。根据英伟达主要芯片规格，可以计算每种芯片的性能密度指标。在新规发布之前，超过旧规性能指标限制的芯片仅为英伟达 A100，但当加入性能密度指标后，新规不仅限制了厂商出略低于性能标准的芯片以规避限制情况，同时针对数据中心芯片与非数据中心芯片进行了不同的限制约定，使更多的英伟达芯片受到禁令限制。

表2：英伟达芯片受限情况

型号	算力性能及性能密度	备注
A100	性能指标 TPP＞4800，性能密度指标＞5.92	受到新规旧规双重限制
H100	性能指标 TPP＞4800，性能密度指标＞5.92	受到新规限制
A800	性能指标 TPP＞4800，性能密度指标＞5.92	受到新规限制
H800	性能指标 TPP＞4800，性能密度指标＞5.92	受到新规限制
L40S	性能指标 TPP＜4800，性能密度处于 1.6~5.92 区间	受到新规限制
RTX4090	性能指标 TPP＜4800，性能密度处于 1.6~5.92 区间	受到新规限制

资料　：芯东西半导体产业媒体，国投证券研究中心

1.2. 人工智能大势所趋，各地政策推进实施

“1+N”政策体系全面推动人工智能产业。2017 年国务院发布《新一代人工智能发展规划》，部委层面陆续出台相关发展规划、实施方案等落地政策，形成“1+N”政策体系，从相关法律法规和伦理规范、人工智能发展支持政策、标准和产权体系、监管和评估体系以及 AI 人才培训等五个角度全面推动人工智能健康快速发展。

同时，各一二线城市均针对 AI 产业制定了产业规模目标和企业数量目标，其中北京市于 2023 年 5 月 30 日发布《北京市加快建设具有全球影响力的人工智能创新策源地实施方案(2023-2025 年)》与《北京市促进通用人工智能创新发展的若干措施》两项重磅政策，以迅速建设具有全球广泛影响力的人工智能创新策源地。

算力发展目标明确，将带动 AI 算力的迅速发展。2023 年 10 月，工业和信息化部、中央网信办、教育部、国家　　、中国人民银行、国务院国资委等六部门联合发布《算力基础设施高质量发展行动计划》，在计算力、运载力、存储力、应用赋能等方面提出了具体目标，以进一步加强算力资源配置，提升国内算力总体水平。智算的快速发展，一方面要求智算中心的建设需要更加合理，要兼顾东西部协同发展和资源的合理利用。另一方面，智能算力更多的采用 AI 芯片，带来更大带宽的网络传输需求，这些都将显著促进 AI 芯片和网络技术的研发创新。

表3：《行动计划》概览

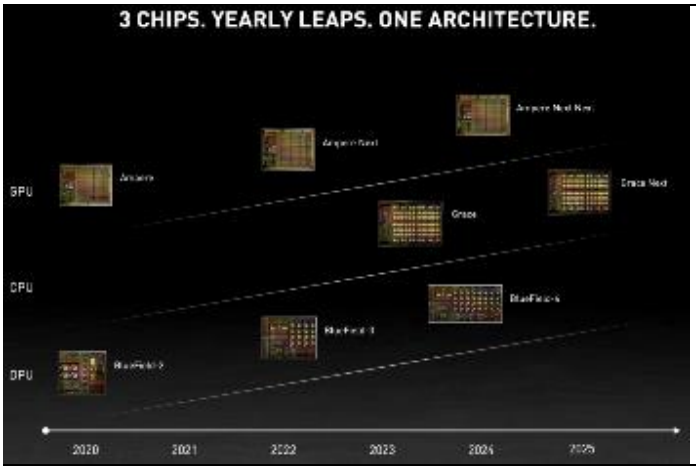
针对方面/地区		政策内容
主要目标（到 2025 年）	计算力	算力规模超过 300EFLOPS，智能算力占比达到 35%，东西部算力平衡协调发展
	运载力	数据中心集群间基本实现不高于理论时延 1.5 倍的直连网络传输，重点应用场所光传送网覆盖率达到 80%
	存储力	存储总量超过 1800EB，先进存储容量占比达到 30%以上，重点行业核心数据、重要数据覆盖率达到 100%
	应用赋能	打造一批算力新业务、新模式、新业态，工业、金融等领域算力渗透率显著提升
重点任务	京津冀、长三角、粤港澳大湾区等	面向重大区域发展战略实施需要有序建设算力设施
	贵州、内蒙古、甘肃、宁夏等节点	推进数据中心集群建设同时，着力提升算力设施利用效率，促进东西部高效互补协同联动

资料　：《算力基础设施高质量发展行动计划》，国投证券研究中心

2. 构建完整的芯片生态系统，“硬件+软件”缺一不可

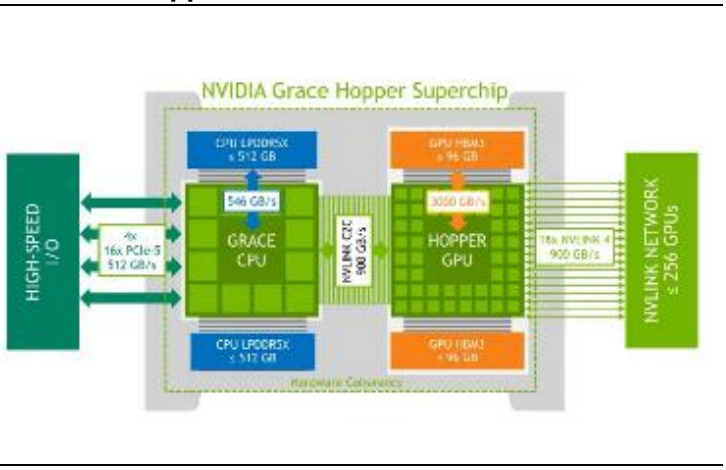
芯片方面，英伟达通过“CPU+GPU+DPU”三芯布局，数据中心正成长为公司最大业务。CPU 作为逻辑处理单元，能更好地处理各种需要快速逻辑判断和并行处理能力的计算任务；GPU 侧重对图像像素进行大规模的数据矩阵运算处理，与 AI 算法的并行结构运算匹配，GPU 在 AI 领域有着先天优势；DPU 则特别适合处理数据中心和网络设备的需求，能有效处理数据包和协议。英伟达通过将 CPU、GPU 和 DPU 集成到同一平台上，可以为客户提供更全面、高效的计算解决方案。公司推出的 Grace Hopper 超级芯片将 Grace 和 Hopper 架构相结合，为加速 AI 和高性能计算（HPC）应用提供 CPU+GPU 相结合的一致内存模型，并在大型服务器上广泛部署。2021 年及以前，游戏业务营收占比最高，但随着 AGI 引爆算力需求，公司数据中心成长极快。根据公司 2022 年年报，其数据中心业务营收约 150 亿美元（占比约 56%），已成为公司最大业务。

图1. “CPU+GPU+DPU”三芯布局



资料　：英伟达 GTC 大会 PPT 截图、国投证券研究中心

图2. Grace Hopper 架构



资料　：公司官网，国投证券研究中心

GPU 领域深耕数十年，长期占据市场领导地位。1999 年英伟达推出的首个 GPU 架构，开创了现代显卡的时代，代表产品是 GeForce 256 显卡，此后其架构经历了多次变革迭代，GPU 计算能力的不断提升，从 2017 到 2022 这五年间，公司先后推出了 Volta、Ampere、Hopper 等针对高性能计算和 AI 训练的架构，并以此为基础发布了 V100、A100、H100 等高端 GPU。通过不断的技术革新，英伟达 GPU 产品向量双精度浮点算力已从 7.8TFLOPS 增至 30TFLOPS。根据 Jon Peddie Research（JPR）报告显示，2023 年 Q1，英伟达 GPU 市场份额达到 84%，Q2 达到 80%，占据市场领导地位。

表4：英伟达 GPU 迭代

架构代号	Fermi	Kepler	Maxwel	Pasca	Volta	Turning	Ampere	Hopper
中文代号	费米	开普勒	麦克斯韦	帕斯卡	伏特	图灵	安培	赫柏
时间	2010	2012	2014	2016	2017	2018	2020	2022
核心参数	16 个 SM，每个 SM 包括 32 Cuda Cores 共计 512 Cuda Cores	15 个 SMx，每个 SMX 包括 192 个单精度+64 个双精度的 Cudacores ；	16 个 SMM，每个 SM 包括 4 个处理块，每个处理块包括 32 个 CUDA 内核+8 个 LD/STUnit+8 个 SFU	Pascal 架构有 GP100、GP102GP100 有 60 个 SM	80 个 SM 每个 SM 里 32 个 FP6464 个 INT3264 个 FP328 个 Tensor core	TU102 核心 72 个 SM，SM 全新设计，每个 SM 里 64 个 INT3264 个 FP328 个 Tensor core	A100 有 108 SMs 每个 SM64 个 FP3264 个 INT3232 个 FP644 个 Tensor core	H100 132 SM 每个 SM128 个 FP3264 个 INT3264 个 FP644 个 Tensor core
特点\优势	首个完整 GPU 计算架构，支持与共享存储结合纯 Cache 层次的 GPU 架构，支持 ECC 的 GPU 架构	游戏性能大幅提首次支持 GPUDirect 技术	相比 Kpler 的每组 SM 单元 192 个减少到了每组 128 个但是每个 SMM 单元拥有更多的逻辑控制电路	每个 SM 包括 64 个 cuda cores32 个 DP cores NVLink 代，双向互联带宽 160GB/sP100 有 56 个 SMHBM	Nvlink 2.0Tensor Core 1.0 满足深度学习和 AI 运算	Tensor Core 2.0RT Core 1.0	Tensor Core 3.0RT Core 2.0Nvlink 3.0 结构稀疏性 MIG 1.0	Tensor Core 4.0Nvlink 4.0 结构稀疏性知阵 MIG 2.0
纳米制程	40/28nm30 亿品体管	28nm71 亿品体管	28nm80 亿品体管	16nm153 亿晶体管	12nm211 亿晶体管	12nm186 亿晶体管	7nm283 亿品体管	4nm800 亿个晶体管
代表型号	Quadro 7000	K80，K40M	M5000M4000	P100GTX 1080P6000	V100TiTan V	T42080TIRTX 5000	A100、A303090	H100

资料 ：英伟达公司官网，国投证券研究中心

2022 年 3 月 GTC 2022 大会上，英伟达正式发布了基于 Hopper 架构的面向数据中心的新一代顶级计算核心 GH100、计算卡 H100。在机器学习及人工智能领域开放产业联盟 MLCommons 公布了最新的 MLPerf 基准评测中，英伟达 H100 Tensor Core GPU 在每次 AI 推理测试中都展现出最高性能。得益于软件优化，该 GPU 的性能比去年 9 月份首次亮相时提高了 54%，，A100 则是英伟达于 2020 年推出的上一代数据中心专用 GPU，但依然是目前 AI 训练的主流芯片产品。根据 New Street Research 的数据，英伟达占据了可用于机器学习的图形处理器市场的 95%。

表5：H100,A100,H800,A800 对比

型号	H100	A100	H800	A800
Target Markets / 目标市场	基于 GH800 图形处理器,SXM5 接口，有助于提高机器学习应用程序的速度，第四代 tensor core， 支持 FPB 和 transformer 引擎	H100 是全球范围内最大的性能出众的 ，拥有革命性的 Transformer 引擎和高度可扩展的 NVIDIA NVLink® 互连技术等突破性功能，可推动庞大的 AI 语言模型、深度推荐系统、基因组学和复杂数字孪生的发展	新引入 FP64，TF32，BF16 Tensor Core，支持 MIG，深度学习/HPC 大规模算例并行加速计算	新引入 FP64，TF32，BF16 Tensor Core，支持 MIG，深度学习/HPC 大规模算例并行加速计算
GPU / 图形处理单元（架构）	GH100（hopper）	GH100（hopper）	GA100（Ampere）	GA100（Ampere）
GPU Cores / CUDA 单元数	16,896	16896	6,912	6,912
VRAM / 显存	80 GB	80GB	80GB	40GB
FP16 Computing（non-Tensor）/ 半精性能 non-Tensor	237.2 TFLOPS	267.6 TFLOPS	77.97 TFLOPS	70 TFLOPS

资料 ：英伟达公司官网，国投证券研究中心

NVLink 是英伟达自研的高速互连技术，解决了多 GPU 并行计算时内存共享和通信的瓶颈问题，能有效提升数据中心的整体运算能力。PCIe 是 Intel 主导的高速串行计算机扩展总线标准，是当前服务器主流的总线解决方案，PCIe 标准迭代周期约为 3 年/代，PCIe 3.0 是目前消费市场的主流选择，4.0 于 2017 年正式推出，自 2021 年下半年开始在数据中心逐步应用，并逐渐从企业级市场下沉到消费市场。目前 Intel/AMD 等主流 CPU 厂商正快速推出 PCIe 5.0 产品，用于 AI 的高性能企业级服务器通常采用 PCIe5.0 接口。NVLink 是英伟达自研的高速接口，可以提供更强大的数据传输能力和更高的吞吐量，能有效缩短数据传输时间，满足当前针对大数据和复杂运算的高带宽需求。随着 NVIDIA GPU 架构的更新和技术的不断发展，NVLink 的版本也在不断演进，以满足不断增长的计算需求和提供更优秀的性能。

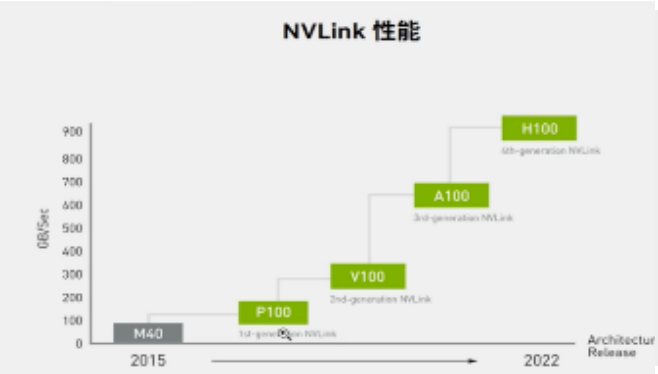
图3. NVLink 与 PCIE 带宽对比

类型	GPU	通道数	双向互联带宽
PCIE互联	A100	PCIE 4.0 x16	2GBx16x2=64GB/s
	H100	PCIE 5.0 x16	4GBx16x2=128GB/s
NVLink互联	A100	每个GPU链路Nvlink x12	25GBx12x2=600GB/s
	H100	每个GPU链路Nvlink x18	25GBx18x2=900GB/s

资料 ：英伟达公司官网，国投证券研究中心

CUDA 架构搭建英伟达软件生态，是英伟达 AI 解决方案的核心优势之一。CUDA（Compute Unified Device Architecture）是由英伟达公司推出的 GPU 加速平台，在推出 CUDA 之前，程序员想要调用 GPU 的运算资源必须先编写大量底层代码，在开发和调试上需要花费大量的时间，而 CUDA 提供了易于使用的编程模型和软件环境，允许开发者使用类似于 C/C++的高层语言进行编程，使得开发者可以借助英伟达的图形处理器（GPU）进行更为高效的并行运算。经过多年优化，目前 CUDA 已成为全球 AI 基础设施，主流的 AI 框架、库、工具都以 CUDA 为基础进行开发。国内第一批大模型厂商使用的基本都是英伟达已经构建完善的 CUDA 生态。即便竞争对手的 GPU 性能的硬件参数上比肩英伟达，如果缺乏 CUDA 的优化，其性能依然无法达到英伟达 GPU 的水平。

图4. NVLink 版本迭代



资料 ：英伟达公司官网，国投证券研究中心

图5. CUDA 架构

资料 ：英伟达公司官网，国投证券研究中心

基于其“芯片+ NVLink + CUDA”的生态系统，英伟达稳坐行业龙头地位，产品全面覆盖 AI 场景。公司 20 多年来始终引领 GPU 行业的发展，将 GPU 的主要应用场景从游戏以及画图等图像显示扩展到了以 AI、云计算等大数据相关的并行计算领域。根据 Jon Peddie Research 发布的 GPU 市场数据统计报告，英伟达 2022 年全年 PCGPU 出货量高达 3034 万块，是 AMD 的

近 4.5 倍；截至 2022 年四季度，在独立 GPU 市场，英伟达占据 84%的市场份额，远超同业竞争公司。

表6：英伟达产品概览

产品线	应用方向	产品
数据中心	AI 训练	Volta 系列、A100 Tensor Core GPU
	AI 推理	Tesla T4、Jetson Xavier NX
	高性能计算	Tesla 系列
游戏	游戏 GPU	GeForce 系列
	游戏笔记本 GPU	RTX 系列
	游戏平板 GPU	Tegra K1
专业可视化	工作站 GPU	Quadro 系列、RTX 系列
	可视化集群	Quadro Virtual Workstation
汽车	汽车 SoC	Drive AGX 系列、Orin 系列
	自动驾驶计算平台	NVDIA DRIVE

资料：公司官网，国投证券研究中心

3. 昇腾软硬件全面布局，构建国产 AI 算力基石

3.1. 昇腾生态包括全栈的 AI 计算基础设施、行业应用及服务

昇腾生态包括昇腾系列处理器、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。

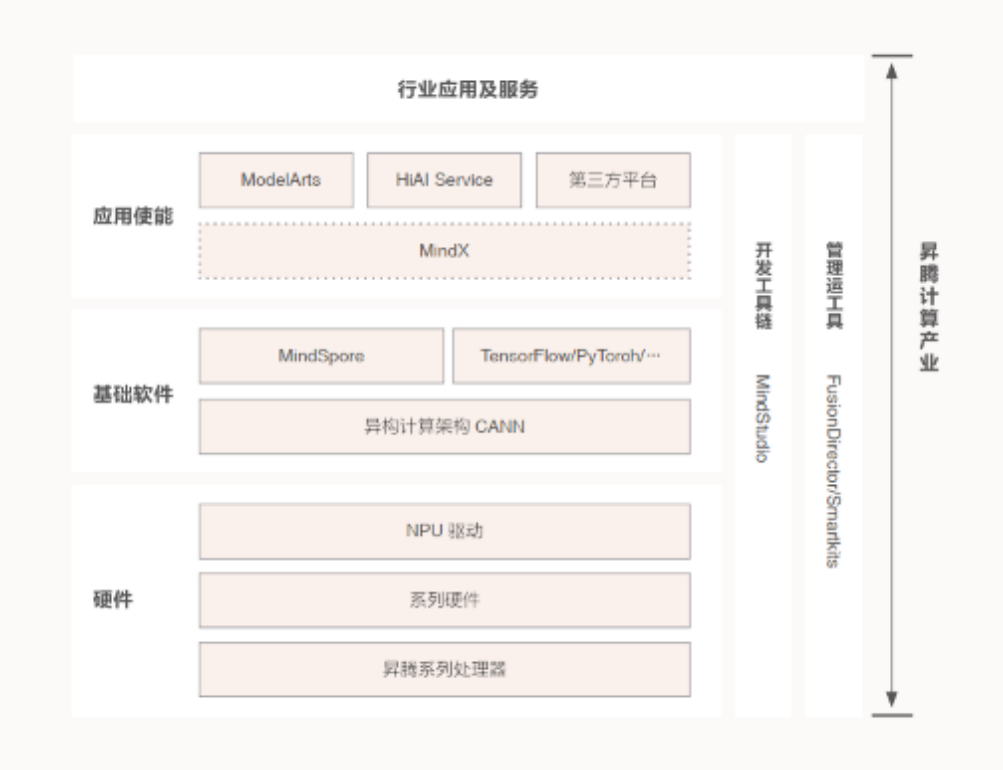
其硬件系统包括：

- 基于　达芬奇内核的昇腾系列处理器等多样化 AI 算力；
- 给予昇腾处理器的系列硬件产品，比如嵌入式模组、板卡、小站、服务器、集群等。

其基础软件体系包括：

- 异构计算架构 CANN 以及对应的驱动、运行时、加速库、编译器、调试调优工具、开发工具链 MindStudio 和各种运维管理工具等；
- AI 计算框架，包括开源的 MindSpore，以及各种业界流行的框架，作为生态的有机组成部分。同时昇腾计算产业支持各种计算框架的对接。

图6. 昇腾生态全景



资料　：昇腾计算产业发展白皮书，国投证券研究中心

基于达芬奇架构的昇腾芯片，运算性能优异，可应用于 AI 训练/推理场景。DaVinci 架构是面向 AI 计算设计的架构，通过独创的 16*16*16 的 3D Cube 设计，每时钟周期可以进行 4096 个 16 位半精度浮点 MAC 计算。同样是完成 4096 次运算，2D 结构需要 64 行*64 列才能计算，3D Cube 只需要 16*16*16 的结构就能算出，因此在核数与频率确定的情况下，每时钟周期进行越多的计算则算力指标越高，而 Davinci 架构的 3D 设计实现了这一点。

从性能指标上来看，昇腾 910 半精度 FP16 的运算性能可达 320TFLOPS，整型 INT8 算力可达 640TOPS，英伟达 A100 的 FP16 运算性能为 312TFLOPS。

表7：昇腾系列处理器算力详情

型号	性能参数	备注
昇腾 310	16TOPS@INT8，8TOPS@FP16	12V 主要应用在边缘计算等低功耗领域
昇腾 910	320TFLOPS@FP16，640TOPS@INT8	2019 年推出，全场景人工智能领域

资料　：昇腾官网，国投证券研究中心

计算代价大幅缩小，功耗水平明显降低。Davinci 架构的 3D 设计以最小的计算代价增加矩阵乘的算力，实现更高的 AI 能效。2018 年 10 月　　联合奥迪展示了 L4 级无人驾驶的路测，汽车上配备了　　的 MDC 车载计算单元，但根据第五届世界互联网大会上前　　公司董事兼　　企业 BG 总裁阎力大披露，**支持 L4 级无人驾驶这样非常复杂的边缘计算场景时，昇腾 310 芯片组仅消耗共计 200 瓦的能耗**，相比英伟达系列芯片均有大幅缩减。

表8：昇腾系列与英伟达系列功耗对比

芯片	设计功耗	备注
昇腾 910	310W	
昇腾 310	8W	
昇腾 310 芯片组	200W	应用于 L4 级无人驾驶等极端复杂边缘计算场景
英伟达 H100	700W	
英伟达 A100	700W	
英伟达 H800	250W	GPU 数目较少，峰值算力低于昇腾 910
英伟达 A800	250W	峰值算力远低于昇腾 910，半精度算力仅为 280TFLOPS

资料：第五届世界互联网大会上前 公司董事兼 企业 BG 总裁阎力大发言，国投证券研究中心

HCCS 是 自研的高速互连接口，可为内核、设备、集群提供系统内存的一致访问，片间带宽最高可达 480Gbps，是业界主流 CPU 互联速率的 2 倍多，HCCS 单个 AI 处理器提供 3 条链路能实现最多 4 个鲲鹏 920 处理器互联和最高 256 个物理核的 NUMA 架构。相比于英伟达 NVLink 与 PCIe 5.0，NVLink 单条链路双向带宽最大为 50GB/s，PCIe 5.0 仅为 4GB/s，HCCS 单条链路双向带宽可以达到 20GB/s，**HCCS 在单一链路的单向/双向互联带宽上比 PCIe 5.0 更具优势，将有效提升多个 AI 处理器协同训练的能力。**

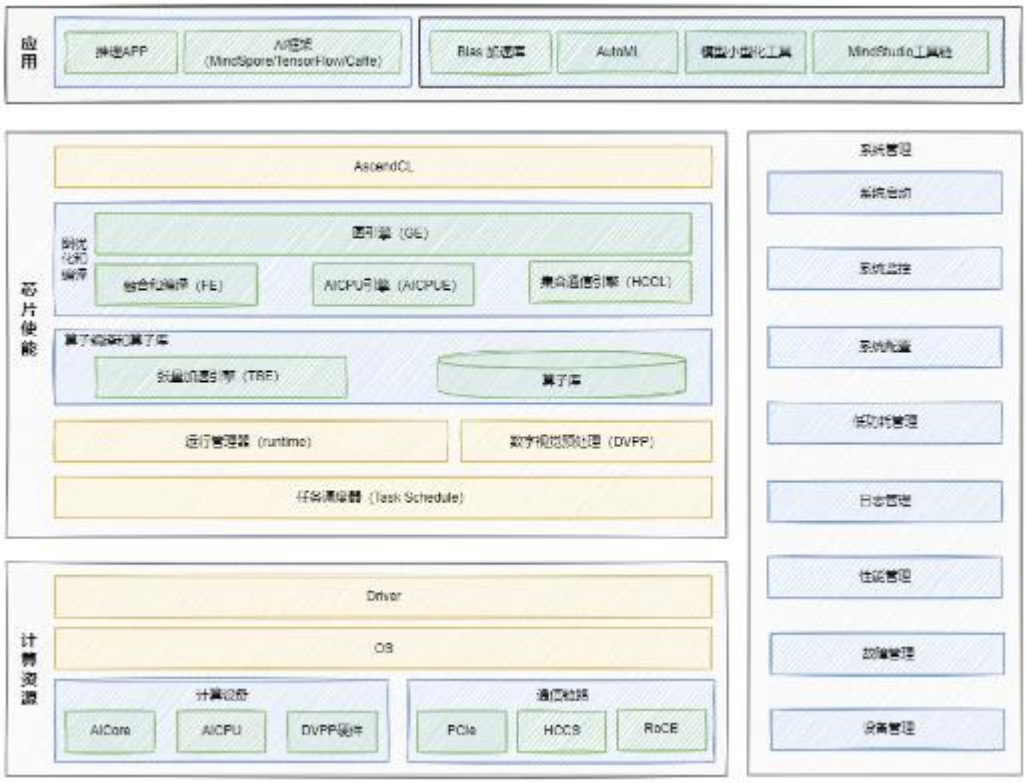
表9：HCCS 接口与 NVLink、PCIe 5.0 对比

类型	链路数	带宽
HCCS	单个 AI 处理器提供 3 条链路	每个 AI 处理器间双向互联带宽 60GB/s（3 条链路）
NVLink	最大提供 18 条链路	单链路双向互联带宽为 50GB/s,18 条链路共计 900GB/s
PCIe 5.0	最多提供 16 条链路	单链路双向互联带宽 8GB/s，16 条链路共计 128GB/s

资料：CSDN，NVIDIA 云数据中心，Atlas 800 训练服务器技术白皮书，国投证券研究中心

CANN 是 针对 AI 场景推出的异构计算架构，通过提供多层次的编程接口，支持用户快速构建基于昇腾平台的 AI 应用和业务。CANN 支持端边云全场景协同，支持超过 10 种设备形态、EMUI、Andriod、openEuler、UOS、Ubuntu、Debian、Suse 等超过 14 种操作系统和多种 AI 计算框架，一套体系支持 CPU、NPU 等架构；

图7. CANN 生态



资料 ： 云-开发者社区-昇腾 AI 基础知识介绍，国投证券研究中心

软件生态建设是 的一大优势。为了帮助 AI 开发者更简单、更高效的开发和使用 AI 技术，推出面向全流程开发工具链 MindStudio。MindStudio 针对算子开发、模型训练、模型推理、应用开发、应用部署的所有全流程工具链进行整合，为开发者提供工程管理、编译、调试、运行、性能分析等全流程开发，提高开发效率。

3.2. 基于“自研芯片+自研接口+自研软件生态”， 推出全系列解决方案

供开发者使用的高性能开发板 Atlas 200 DK，Atlas 200 DK 开发者套件（型号 3000）是以 Atlas 200 AI 加速模块（型号 3000）为核心的开发者板形态终端类产品（其中 Atlas 200 AI 加速模块是高性能 AI 计算模块，集成了昇腾 310 AI 处理器，芯片内置 2 个 AI core，可支持 128 位宽的 LPDDR4X，最大算力为 22TOPS）。

表10： Atlas 200 DK 开发者套件硬件规格

特征	规格
AI 处理器（昇腾 310AI 处理器）	2 个 DaVinci AI Core（达芬奇内核） 8 个 A55 Arm Core(最大主频 1.6GHz)
AI 算力	半精度（FP16）：4/8/11 TFLOPS 整数精度（INT8）：8/16/22 TOPS
内存	类型：LPDDR4X 位宽：128bit/64bit 容量：8GB/4GB 速率：3200Mbps 支持 ECC（Error Correction Code）
接口	千兆 网接口：1 个 GE RJ-45 接口 USB 接口：1 个 USB3.0 Type C 接口 1 个 40pin IO 连接器，2 个 MIPI 连接器，2 个板载麦克风

资料 ： 昇腾产品技术白皮书，国投证券研究中心

训练卡 Atlas 300 T。Atlas 300 T 训练卡（型号 9000）可以配合服务器为数据中心提供 AI 加速卡，单卡最高可提供 220 TFLOPS FP16 算力。产品具有强算力、高度集成、高速带宽等特点，可满足大量人工智能训练以及高性能计算领域的算力需求。

表11：Atlas 300 T 训练卡（型号 9000）硬件规格

特征	规格
AI 处理器（昇腾 910 AI 处理器）	30 个达芬奇 AI Core+ 16 个 TaiShan 核
内存规格	32GB HBM 16GB DDR4 2400Mbps 支持 ECC（Error Connection Code）
AI 算力	半精度（FP16）：最大 220 TFLOPS 整数精度（INT8）：最大 440 TOPS
PCIe 接口	PCIe *16 Gen4.0
网络	1*100GE QSFP-DD 接口，支持 RoCE 协议

资料 ： 昇腾产品技术白皮书，国投证券研究中心

推理卡 Atlas 300 I 。Atlas 300I 推理卡采用 4 个昇腾 310AI 处理器的 PCIe HHHL 卡，实现快速高效的推理计算、图像识别及视频处理等工作，支持多种规格的 H.264、H.265 视频编解码。

表12：Atlas 300I 推理卡（型号 3000）硬件规格

特征	规格
AI 处理器（4 个昇腾 310AI 处理器，每个处理器包含：）	2 个 DaVinci AI Core（达芬奇内核） 8 个 A55 Arm Core(最大主频 1.6GHz)
AI 算力	半精度（FP16）：44 TFLOPS 整数精度（INT8）：88 TOPS
内存	类型：LPDDR4X 位宽：512bit 容量：32GB 速率：3200Mbps 总带宽：204.8GByte/s 支持 ECC（Error Correction Code）
PCIe 接口	PCIe3.0*8，兼容 PCIe2.0/PCIe1.0

数据：昇腾产品技术白皮书，国投证券研究中心

Atlas 800 训练服务器，8 颗昇腾算力芯片+4 颗鲲鹏 CPU。Atlas 800 训练服务器（型号：9000）是基于 鲲鹏+昇腾处理器的 AI 训练服务器，具有超强算力密度、超高能效与高速网络带宽等特点。该服务器广泛应用于深度学习模型开发和训练，适用于智慧城市、智慧医疗、天文探索、石油勘探等需要大算力的行业领域。

表13：Atlas 800 训练服务器规格

特征	规格
CPU	4*鲲鹏 920 处理器
AI 处理器	8*昇腾处理器
内存规格	最多 32 个 DDR4 内存插槽 内存最高速率 3200MT/s 单根内存条容量支持 16/32/64GB
PCIe 拓展	最多支持 2 个 PCIe 4.0 拓展插槽
功耗	最大功耗 5.6KW

资料：昇腾产品技术白皮书，国投证券研究中心

Atlas 800 推理服务器，8 颗昇腾推理卡+2 颗鲲鹏 CPU。Atlas 800 推理服务器（型号：3000）是基于昇腾处理器的推理服务器，最大可支持 8 个 Atlas 300I 推理卡，提供强大的实时推理能力，广泛应用于中心侧 AI 推理场景。

表14：Atlas 800 推理服务器硬件规格

特征	规格
CPU	2*鲲鹏 920
AI 加速卡	最大支持 8 个 Atlas 300I 推理卡
内存规格	鲲鹏 920 7260/5250：最多 32 个 DDR4 内存插槽
PCIe 扩展槽位	鲲鹏 920 5220/3210：最多 16 个 DDR4 内存插槽 内存设计最大速率 2933MT/s 单根内存条支持 8/16/32/64/128GB 最多支持 9 个 PCIe4.0 PCIe 接口
端口	（1）前面板提供 2 个 USB 3.0 端口、1 个 DB15 VGA 端口 （2）后面板提供 2 个 USB3.0 端口，1 个 DB15 VGA 端口、1 个 RJ45 管理串口、1 个 RJ45 管理网口
显卡	显卡芯片集成在 iBMC 管理芯片中，提供 32MB 显存，支持最高 60Hz 频率下 16M 色彩的最大分辨率是 1920*1080 像素

资料： 昇腾产品技术白皮书，国投证券研究中心

由数千颗昇腾处理器构成的 Atlas 900 AI 集群。Atlas 900 AI 集群由数千颗昇腾处理器构成，整合 HCCS、 PCIe 4.0 和 100G RoCE 三种高速接口。其总算力达到 256P～1024P FLOPS @FP16，相当于 50 万台 PC 的计算能力。它可以在 60 秒完成基于 Resnet-50 模型训练，比第 2 名快 15%，这可以让使用者更快的进行 AI 训练，高效地推进预测天气、勘探石油、自动驾驶等等商用进程。

图8. Atlas 900 AI 集群

资料： 昇腾开发者论坛，国投证券研究中心

4.科技巨头纷纷布局算力芯片，AI 浪潮势不可挡

4.1.AMD 的“MI 系列+Infinity Fabric+ROCm 平台”，成为英伟达全球范围内最强劲的对手

Radeon Instinct 系列是 AMD 专为数据中心和企业市场推出的 GPU 解决方案，旨在支持深度学习、高性能计算和科学研究等。从 2017 年发布 Radeon Instinct MI6，到如今更新至 Radeon Instinct MI300 系列，采用高性能的 GCN 或 RDNA 架构，支持大规模的并行计算和机器学习任务。同时支持 ROCm（Radeon Open Compute）平台，以提供开发和部署机器学习模型的工具和库。

表15：AMD Redeon Instinct 系列产品迭代

型号	发布日期	GPU 架构	光刻	显存规格	峰值功耗	计算单元	峰值性能（FP16）
MI6	06/2017	Polaris	14nm FinFET	16GB（GDDR5）	150W		5.73 TFLOPS
MI8	06/2017	GCN 3rd Gen	28nm	4GB（HBM）	175W		8.19 TFLOPS
MI25	06/2017	Vega	14nm FinFET	16GB（HBM2）	300W	64	24.6 TFLOPS
MI60	11/18/2018	Vega20	TSMC 7nm FinFET	32GB（HBM2）	300W	64	29.5 TFLOPS
MI50（16GB/32GB）	11/18/2018	Vega20	TSMC 7nm FinFET	16GB/32GB（HBM2）	300W	60	26.5 TFLOPS
MI100	11/16/2020	CDNA	TSMC 7nm FinFET	32GB（HBM2）	300W	120	184.6 TFLOPS
MI250	11/08/2021	CDNA2	TSMC 6nm FinFET	128GB（HBM2e）	560W	208	362.1 TFLOPS
MI250X	11/08/2022 1	CDNA2	TSMC 6nm FinFET	128GB（HBM2e）	560W	220	383 TFLOPS
MI210	03/22/2022	CDNA2	TSMC 6nm FinFET	64GB（HBM2e）	300W	104	181 TFLOPS
MI300X	06/13/2023	CDNA3	TSMC 5nm FinFET	192GB（HBM3）	800W	320	2615 TFLOPS

资料：AMD 官网，国投证券研究中心

2023 年 AMD 公司推出 Radeon Instinct MI300 系列，正式迈进“百亿亿级计算”时代。AMD Instinct MI300 系列基于 AMD CDNA 3 架构打造，包括 AMD Instinct MI300A APU（创新的 AI 和 HPC 工作负载专用 APU）和 AMD Instinct MI300X GPU，可为广泛的AI和HPC工作负载提供领先的应用程序性能。随着 AI 工作负载的扩展，AMD Instinct MI300X 提供了采用 UBB 业界标准 OCP 平台设计的普适性解决方案，支持客户将 8 个 GPU 整合为一个性能主导型节点，并且具有全互联式点对点环形设计，单一平台内的 HBM3 显存总计可达到 1.5 TB——提供足以应对各类 AI 或 HPC 工作负载部署的性能密集型解决方案。

2023 年 6 月，AMD 首席执行官苏姿丰（Lisa Su）在旧金山举行的发布会上表示，MI300X 提供的 HBM 密度最高是英伟达 AI 芯片 H100 的 2.4 倍，其 HBM 带宽最高是 H100 的 1.6 倍。MI300X 是针对 LLM 的优化版，拥有 192GB 的 HBM3 内存、5.2TB/秒的带宽和 896GB/秒的 Infinity Fabric 带宽。AMD 将 1530 亿个晶体管集成在共 12 个 5 纳米的小芯片中。

Infinity Fabric 是 AMD 的高速接口技术，用于连接 CPU 和 GPU 内部的不同部分，以及连接不同的 CPU 和 GPU，理论峰值 P2P I/O 带宽最高可达 896 GB/s，与 NV Link 旗鼓相当。多达 8 个 Infinity Fabric 链接将 AMD Instinct MI300X 与节点中的第三代 EPYC 处理器和其他 GPU 相连，以实现统一的 CPU 内存/GPU 显存一致性和系统吞吐量最大化，通过的强大性能使 CPU 代码更简化。