计算机行业深度报告

国产 AI 算力行业报告：浪潮汹涌，势不可挡

增持（维持）

2024 年 03 月 26 日

证券分析师 王紫敬
执业证书：S0600521080005
021-60199781
wangzj@dwzq.com.cn
证券分析师 王世杰
执业证书：S0600523080004
wangshijie@dwzq.com.cn

投资要点

- **海外应用、算力和模型相互演进，AI 浪潮滚滚而来：**2024 年 2-3 月，OpenAI 发布 Sora，Anthropic 发布了新一代 AI 大模型系列——Claude 3，马斯克开源大模型 Grok-1，英伟达在 GTC 大会上推出新一代 GPU GB200，全球 AI 产业发展速度逐步加快。
- **国内模型、应用不断突破，算力需求逐步放大：**2024 年 3 月 18 日，Kimi 上下文长度提升到 200 万字，访问量大幅提升，算力告急。3 月 23 日，阶跃星辰发布了万亿参数大模型预览版，标志着国产 AI 大模型取得了巨大进步。国产 AI 大模型正在不断迭代，对算力需求会不断提升。
- **国内 AI 芯片需求旺盛：**在英伟达 GTC 大会上，黄仁勋讲到，如果要训练一个 1.8 万亿参数量的 GPT 模型，需要 8000 张 H100，用时 90 天。我们测算如果中国有十家大模型公司要达到 GPT-4 水平，则需要 8 万张 H100 GPU。我们预计，推理算力需求将是训练的数倍，高达几十万张 H100。
- **政策加持叠加海外制裁，国产 AI 芯片需求会逐步加快：**虽然国产 AI 芯片在单卡性能、生态和集群效率上与海外产品仍有一定差距，但改进速度较快，已经形成万卡集群，并在科大讯飞、部分互联网大厂用于 AI 大模型训练。3 月 22 日，上海政策要求，到 2025 年，上海市新建智算中心国产算力芯片使用占比超过 50%。
- **国产 AI 芯片中，昇腾一马当先，各家竞相发展：**华为昇腾是国产 AI 芯片龙头，根据财联社报道，2022 年昇腾占据国内智算中心约 79%的市场份额。海光信息、寒武纪、景嘉微等公司国产 AI 芯片产品均已有下游客户测试使用，后续有望迎来放量。
- **算力产业蓬勃发展，多个细分方向迎来机会：算力租赁。**AI 算力租赁刚刚兴起，参与方众多，格局还比较分散。AI 算力租赁目前的核心竞争力是谁能拿到优质计算卡。**算力液冷。**3 月 19 日，GTC 大会英伟达提出 GB200 使用液冷方案。液冷技术壁垒不高，行业壁垒较高。根据我们测算，2025 年及以后存量服务器改造为冷板式液冷市场空间为 832 亿元；假设 2027 年新增 AI 服务器全部采用冷板式液冷，市场规模为 260 亿元。**全国一体化算力网。**算力调度类似于电力调度。央国企有望在算力调度中大有作为。2025 年，我们测算悲观、中性和乐观情况下，对应算力调度市场规模为 444、710、887 亿元。**央企 AI。**2 月 19 日，国资委明确要求中央企业要把发展人工智能放在全局工作中统筹谋划，深入推进产业焕新，加快布局和发展人工智能产业。
- **投资建议：**不论国内还是海外，大模型和应用都在不断迭代和发展，算力需求增加的确定性会越来越强。但由于海外制裁和国家政策支持，算力国产化比例会逐渐提高。同时，算力的新技术、新方向也会逐步发展起来。
- **相关标的：国产算力：**华为系：神州数码、软通动力、高新发展、拓维信息等。海光系：海光信息、中科曙光。其他：寒武纪、景嘉微等。**算力一体化：**广电运通、博睿数据、思特奇、恒为科技、美利云等。**算力租赁：**云赛智联、润泽科技、利通电子、润建股份、迈信林等。**算力液冷：**英维克、网宿科技、高澜股份、精研科技等。**央企 AI：**国投智能、新华网等。其他：九联科技。
- **风险提示：**政策支持不及预期；技术发展不及预期；AI 发展不及预期。

行业走势

相关研究

《AI 算力不断迭代，液冷大势所趋》

2024-03-11

《数据要素的报台账时刻：关注新政策方向》

2024-02-27

请务必阅读正文之后的免责声明部分

内容目录

请务必阅读正文之后的免责声明部分

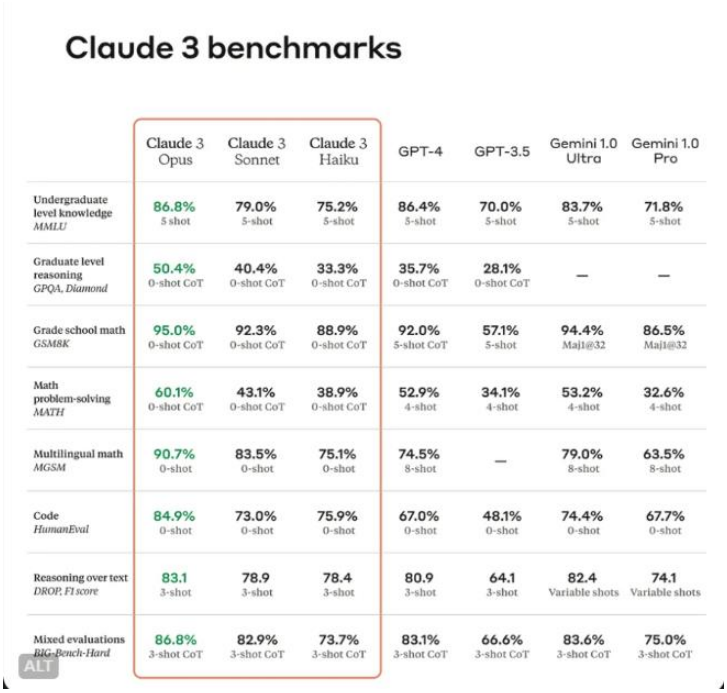
图表目录

1. 海外：模型、应用和算力相互推进

2 月 16 日，OpenAI 发布了首个文生视频模型 Sora。Sora 可以直接输出长达 60 秒的视频，并且包含高度细致的背景、复杂的多角度镜头，以及富有情感的多个角色。

3 月 4 日，Anthropic 发布了新一代 AI 大模型系列——Claude 3。该系列包含三个模型，按能力由弱到强排列分别是 Claude 3 Haiku、Claude 3 Sonnet 和 Claude 3 Opus。其中，能力最强的 Opus 在多项基准测试中得分都超过了 GPT-4 和 Gemini 1.0 Ultra，在数学、编程、多语言理解、视觉等多个维度树立了新的行业基准。Claude 首次带来了对多模态能力的支持（Opus 版本的 MMMU 得分为 59.4%，超过 GPT-4V，与 Gemini 1.0 Ultra 持平）。

图1：Claude 3 benchmarks



数据来源：Anthropic，东吴证券研究所

3 月 18 日，马斯克开源大模型 Grok-1。马斯克旗下 AI 初创公司 xAI 宣布，其研发的大模型 Grok-1 正式对外开源开放，用户可直接通过磁链下载基本模型权重和网络架构信息。xAI 表示，Grok-1 是一个由 xAI 2023 年 10 月使用基于 JAX 和 Rust 的自定义训练堆栈、从头开始训练的 3140 亿参数的混合专家（MOE）模型，远超 OpenAI 的 GPT 模型。

在 CEO 奥尔特曼的带领下，OpenAI 或许有望在今年夏季推出 GPT-5。

3 月 23 日，媒体援引知情人士透露，OpenAI 计划下周在美国洛杉矶与好莱坞的影视公司和媒体高管会面。OpenAI 希望与好莱坞合作，并鼓励电影制作人将 OpenAI 最新 AI 视频生成工具 Sora 应用到电影制作中，从而拓展 OpenAI 在娱乐行业的影响力。

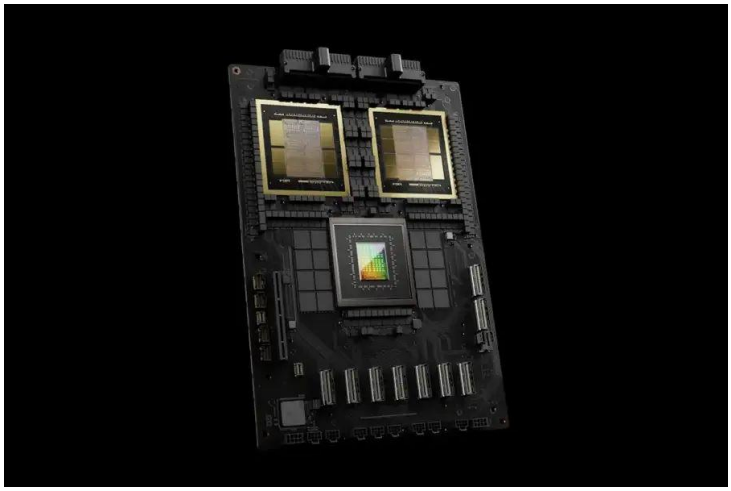
请务必阅读正文之后的免责声明部分

3 月 19 日，英伟达 GTC 大会上，英伟达发布新的 B200 GPU，以及将两个 B200 与单个 Grace CPU 相结合的 GB200。

全新 B200 GPU 拥有 2080 亿个晶体管，采用台积电 4NP 工艺节点，提供高达 20 petaflops FP4 的算力。与 H100 相比，B200 的晶体管数量是其（800 亿）2 倍多。而单个 H100 最多提供 4 petaflops 算力，直接实现了 5 倍性能提升。

而 GB200 是将 2 个 Blackwell GPU 和 1 个 Grace CPU 结合在一起，能够为 LLM 推理工作负载提供 30 倍性能，同时还可以大大提高效率。

图2：GB200 超级芯片



数据来源：英伟达，东吴证券研究所

计算能力不断提升。过去，训练一个 1.8 万亿参数的模型，需要 8000 个 Hopper GPU 和 15MW 的电力。如今，2000 个 Blackwell GPU 就能完成这项工作，耗电量仅为 4MW。在 GPT-3（1750 亿参数）大模型基准测试中，GB200 的性能是 H100 的 7 倍，训练速度是 H100 的 4 倍。

2. 国内模型逐步追赶，提升算力需求

Kimi 逐渐走红。月之暗面 Kimi 智能助手 2023 年 10 月初次亮相时，凭借约 20 万汉字的无损上下文能力，帮助用户解锁了专业学术论文的翻译和理解、辅助分析法律问题、一次性整理几十张发票、快速理解 API 开发文档等，获得了良好的用户口碑和用户量的快速增长。

2024 年 3 月 18 日，Kimi 智能助手在长上下文窗口技术上再次取得突破，无损上下文长度提升了一个数量级到 200 万字。

过去要 10000 小时才能成为专家的领域，现在只需要 10 分钟，Kimi 就能接近任何一个新领域的初级专家水平。用户可以跟 Kimi 探讨这个领域的问题，让 Kimi 帮助自己练习专业技能，或者启发新的想法。有了支持 200 万字无损上下文的 Kimi，快速学习任何一个新领域都会变得更加轻松。

请务必阅读正文之后的免责声明部分

访问量提升，kimi 算力告急。3 月 21 日下午，大模型应用 Kimi 的 APP 和小程序均显示无法正常使用，其母公司月之暗面针对网站异常情况发布说明：从 3 月 20 日 9 点 30 分开始，观测到 Kimi 的系统流量持续异常增高，流量增加的趋势远超对资源的预期规划。这导致了从 20 日 10 点开始，有较多的 SaaS 客户持续的体验到 429:engine is overloaded 的异常问题，并对此表示深表抱歉。

2024 年 3 月 23 日，阶跃星辰发布 Step 系列通用大模型。产品包括 Step-1 千亿参数语言大模型、Step-1V 千亿参数多模态大模型，以及 Step-2 万亿参数 MoE 语言大模型的预览版，提供 API 接口给部分合作伙伴试用。

相比于 GPT-3.5 是一个千亿参数模型，GPT-4 是拥有万亿规模参数，国内大模型厂商如果想追赶，需要各个维度要求都上一个台阶。

阶跃星辰发布了万亿参数大模型预览版，标志着国产 AI 大模型取得了巨大进步。

国产 AI 大模型正在不断迭代，对算力需求会不断提升。

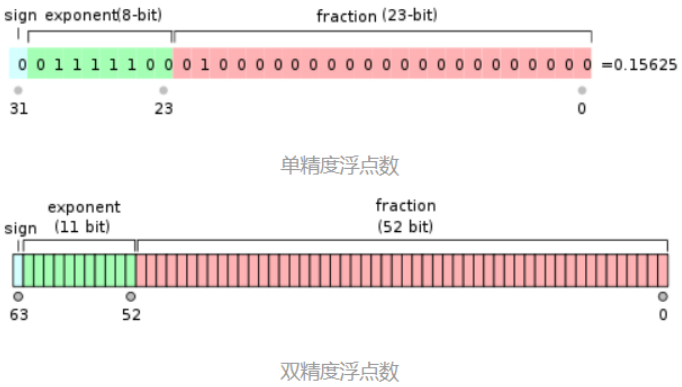
3. 国内算力产业现状盘点

3.1. 算力有哪些核心指标?

算力芯片的主要参数指标为算力浮点数，显存，显存带宽，功耗和互连技术等。

算力浮点数：算力最基本的计量单位是 FLOPS，英文 Floating-point Operations Per Second，即每秒执行的浮点运算次数。算力可分为双精度(FP64)，单精度(FP32)，半精度(FP16)和 INT8。FP64 计算多用于对计算精确度要求较高的场景，例如科学计算、物理仿真等；FP32 计算多用于大模型训练等场景；FP16 和 INT8 多用于模型推理等对精度要求较低的场景。

图3：GPU 算力浮点数图示



数据来源：CSDN，东吴证券研究所

GPU 显存：显存用于存放模型，数据显存越大，所能运行的网络也就越大。

请务必阅读正文之后的免责声明部分

在预训练阶段，大模型通常选择较大规模的数据集获取泛化能力，因此需要较大的批次等来保证模型的训练强大。而模型的权重也是从头开始计算，因此通常也会选择高精度（如 32 位浮点数）进行训练。需要消耗大量的 GPU 显存资源。

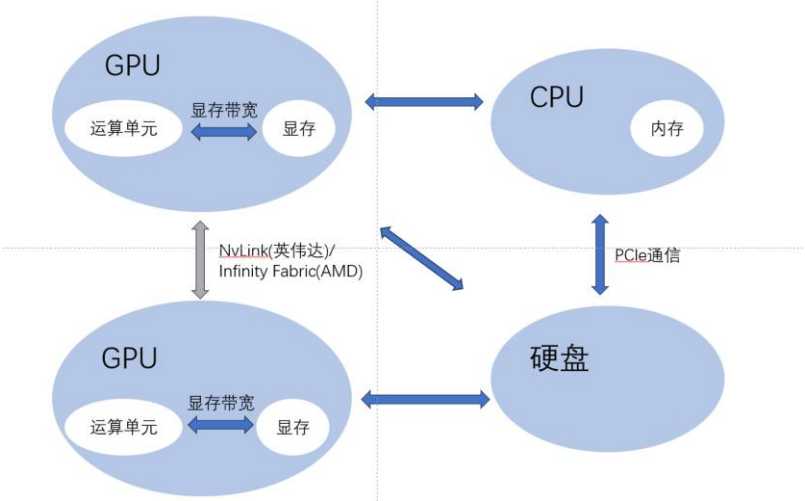
在微调阶段，通常会冻结大部分参数，只训练小部分参数。同时，也会选择非常多的优化技术和较少的高质量数据集来提高微调效果，此时，由于模型已经在预训练阶段进行了大量的训练，微调时的数值误差对模型的影响通常较小。也常常选择 16 位精度训练。因此通常比预训练阶段消耗更低的显存资源。

在推理阶段，通常只是将一个输入数据经过模型的前向计算得到结果即可，因此需要最少的显存即可运行。

显存带宽：是运算单元和显存之间的通信速率，越大越好。

互连技术：一般用于显存之间的通信，分布式训练，无论是模型并行还是数据并行，GPU 之间都需要快速通信，不然就是性能的瓶颈。

图4：关键参数关系示意图



数据来源：东吴证券研究所绘制

3.2. 国产算力和海外的差距

从单芯片能力看，训练产品与英伟达仍有 1-2 代硬件差距。根据科大讯飞，华为昇腾 910B 能力已经基本做到可对标英伟达 A100。推理产品距离海外差距相对较小。

图5：主流国内外 AI 芯片性能对比

公司	型号	场景	生产工艺	INT8算力	FP16算力	FP32算力	FP64算力	最大功耗 TDP	显存带宽 GB/s
华为昇腾	昇腾310	推理	12nm FFC	16 TOPS	8 TOPS			8W	
	昇腾910	训练	N7+	640 TOPS	320 TFLOPS			310W	
寒武纪	MLU370-S4	推理	7nm	192 TOPS	96 TOPS	18 TFLOPS		75W	307.2 GB/s
	MLU370-X4	训练+推理	7nm	256 TOPS	96 TFLOPS	24 TFLOPS		150W	307.2 GB/s
	MLU370-X8	训练+推理	7nm	256 TOPS	96 TFLOPS	24 TFLOPS		250W	614.4 GB/s
	MLU290-M5	训练	7nm	512 TOPS	TOPS (INT16)	OPS (CINT32)		350W	1228 GB/s
	MLU270-S4	推理		128 TOPS	TOPS (INT16)			70w	102 GB/s
	MLU270-F4	推理		128 TOPS	TOPS (INT16)			150w	102 GB/s
景嘉微	JM9100					512G FLOPS		5-15W	25.6GB/s
	JM92系列					1.2T FLOPS		15-30W	128GB/s
	M9系列					1.5T Flops		<30W	128GB/s
海光	深算一号		7nm FinFET					350 W	1024 GB/s
燧原科技	云燧T20	训练		256 TOPS	128 TFLOPS	32 TFLOPS		300W	1.6TB/s
	云燧T21	训练		256 TOPS	128 TFLOPS	32 TFLOPS		300W	1.6TB/s
	云燧i21	推理		256 TOPS	128 TFLOPS	32 TFLOPS		150W	819 GB/s
英伟达	V100 PCIe	训练+推理	12nm FFN			14 TFLOPS	7 TFLOPS	250W	900 GB/s
	V100 SXM2	训练+推理				15.7 TFLOPS	7.8 TFLOPS	300W	900 GB/s
	V100S PCIe	训练+推理				16.4 TFLOPS	8.2 TFLOPS	250W	1134 GB/s
	A100 80GB P(	训练+推理	7nm		312 TFLOPS	19.5 TFLOPS	9.7 TFLOPS	300W	1935 GB/s
	A100 80GB S)	训练+推理			312 TFLOPS	19.5 TFLOPS	9.7 TFLOPS	400W*	2039 GB/s
	H100 SXM	训练+推理	4nm		1979 TFLOPS	67 teraFLOPS	34 teraFLOPS	700W	3.35TB/s
	H100 PCIe	训练+推理			1513 TFLOPS	51 teraFLOPS	26 teraFLOPS	300-350W	2TB/s
AMD	MI250	训练	TSMC 6nm Fi	362.1 TOPs	362.1 TFLOPs	45.3 TFLOPs	45.3 TFLOPs		100 GB/s
	MI250X	训练	TSMC 6nm Fi	383 TOPs	383 TFLOPs	47.9 TFLOPs	47.9 TFLOPs		100 GB/s

*400W TDP（适用于标准配置）。HGX A100-80 GB 自定义散热解决方案 (CTS) SKU 可支持高达 500W 的 TDP

数据来源：公司官网，东吴证券研究所

从片间互联看，片间和系统间互联能力较弱。国产 AI 芯片以免费 CCIX 为主，生态不完整，缺少实用案例，无 NV-Link 类似的协议。大规模部署稳定性和规模性距离海外仍有较大差距。

从生态看，大模型多数需要在专有框架下才能发挥性能，软件生态差距明显，移植灵活性，产品易用性与客户预期差距较大。客户如果使用国产 AI 芯片，需要额外付出成本。

从研发能力看，产品研发能力（设计与制程），核心 IP（HBM，接口等）等不足，阻碍了硬件的性能提升。

3.3. 国产化和生态抉择

海外制裁后，AI 芯片国产化诉求加大。主要系供应链安全和政策强制要求。

2024 年 3 月 22 日，上海市通信管理局等 11 个部门联合印发《上海市智能算力基础设施高质量发展 “算力浦江”智算行动实施方案（2024-2025 年）》。**到 2025 年，上海市市新建智算中心国产算力芯片使用占比超过 50%，**国产存储使用占比超过 50%，服务具有国际影响力的通用及垂直行业大模型设计应用企业超过 10 家。

但国产 AI 芯片由于生态、稳定性、算力等问题，目前较多用于推理环节，少数用

请务必阅读正文之后的免责声明部分

于训练。如用于训练，则需花费较多人员进行技术服务，额外投入资源较大。

华为与讯飞构建昇腾万卡集群。2023 年 10 月 24 日，科大讯飞携手华为，宣布首个支撑万亿参数大模型训练的万卡国产算力平台“飞星一号”正式启用。1 月 30 日，讯飞星火步履不停，基于“飞星一号”，启动了对标 GPT-4 的更大参数规模的大模型训练。

“飞星一号“是科大讯飞和华为联合发布基于昇腾生态的国内首个可以训练万亿浮点参数大模型的大规模算力平台。也是国内首个已经投产使用的全国产大模型训练集群，采用昇腾 AI 硬件训练服务器和大容量交换机构建参数面无损 ROCE 组网，配置高空间的全闪和混闪并行文件系统，可支撑万亿参数大模型高速训练。

3.4. 国内算力厂商竞争要素

在中国市场，算力行业的核心竞争要素为供应链安全、服务能力、政府关系、资金、技术、人才等。

供应链安全。受美国制裁影响，众多算力芯片厂商芯片供应链出现问题。如果能够解决供应链问题，持续为客户供应芯片，将是一大核心竞争力。

服务能力。AI 算力集群的构建后续的运维需要强大的服务支持，对于生态基础较弱的国产芯片厂商要求更高。

政府关系。国产 AI 芯片的采购一大驱动为政策支持，具有良好的政府关系和客户渠道，可以打开市场空间。

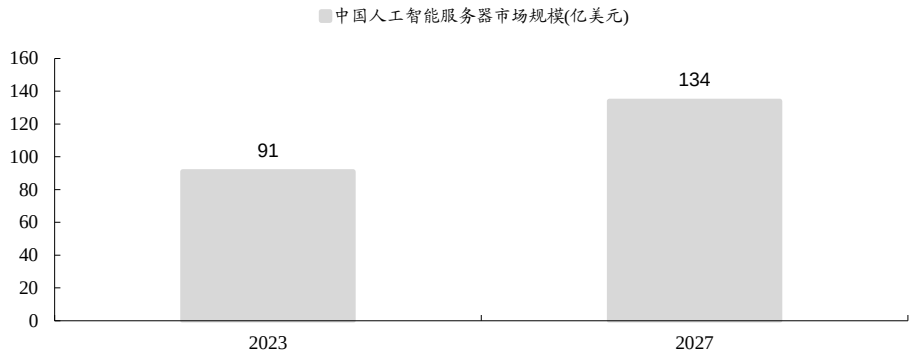
资金、技术和人才。AI 芯片的研发和突破需要大量的资源投入，我们看好具备强大资金、技术和人才储备的公司。

3.5. 国内 AI 算力市场空间

IDC 报告预计，2023 年中国人工智能服务器市场规模将达 91 亿美元，同比增长 82.5%，2027 年将达到 134 亿美元，2022-2027 年年复合增长率达 21.8%。

图6：中国 AI 服务器市场规模

请务必阅读正文之后的免责声明部分



数据来源：IDC，东吴证券研究所

算力需求市场空间巨大。在英伟达 GTC 大会上，黄仁勋讲到，如果要训练一个 1.8 万亿参数量的 GPT 模型，需要 8000 张 Hopper GPU，消耗 15 兆瓦的电力，连续跑上 90 天。如果中国有十家大模型公司，则需要 8 万张 H100 GPU。我们预计，推理算力需求将是训练的数倍，高达几十万张 H100。随着模型继续迭代，算力需求只会越来越大。

随着国产化率逐步提升，我们预计 AI 芯片逐步成为国内芯片的主要组成。

4. 国内供给端：昇腾一马当先，各家竞相发展

北京商报对华为公司董事长梁华的主题演讲的分享中提到，昇腾已经在华为云和 28 个城市的智能算力中心大规模部署，根据财联社报道，2022 年昇腾占据国内智算中心约 79%的市场份额。

4.1. 昇腾计算产业链

华为主打 AI 芯片产品有 310 和 910B。310 偏推理，当前主打产品为 910B，拥有 FP32 和 FP16 两种精度算力，可以满足大模型训练需求。910B 单卡和单台服务器性能对标 A800/A100。

昇腾计算产业是基于昇腾 AI 芯片和基础软件构建的全栈 AI 计算基础设施、行业应用及服务，能为客户提供 AI 全家桶服务。主要包括昇腾 AI 芯片、系列硬件、CANN、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。

硬件系统：基于华为达芬奇内核的昇腾系列 AI 芯片；基于昇腾 AI 芯片的系列硬件产品，比如嵌入式模组、板卡、小站、服务器、集群等。

软件系统：

请务必阅读正文之后的免责声明部分

异构计算架构 CANN 以及对应的调试调优工具、开发工具链 MindStudio 和各种运维管理工具等。

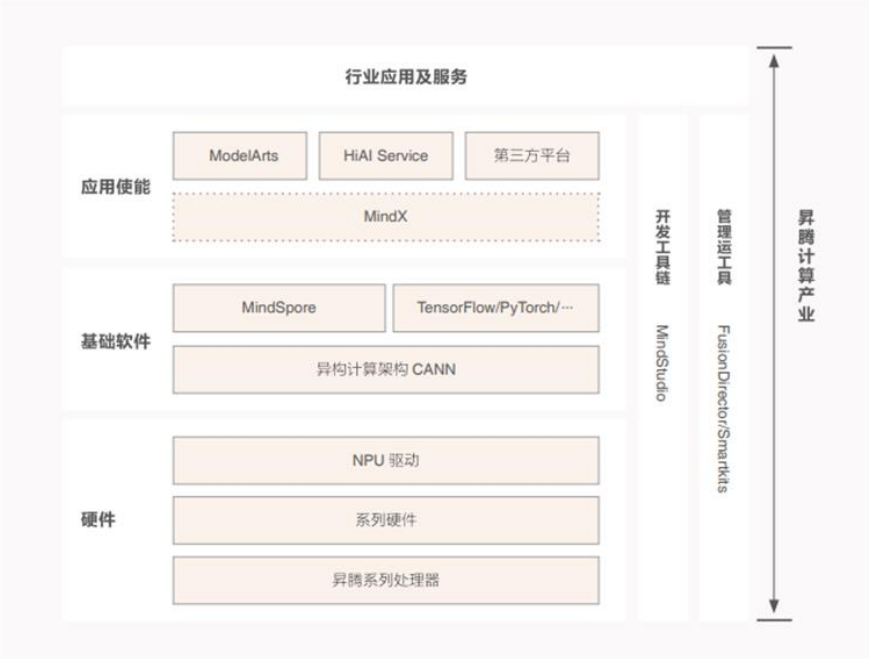
AI 计算框架包括开源的 MindSpore,以及各种业界流行的框架。

昇思 MindSpore AI 计算架构位居 AI 框架第一梯队。

下游应用：昇腾应用使能 MindX，可以支持上层的 ModelArts 和 HiAI 等应用使能服务。

行业应用是面向千行百业的场景应用软件和服务，如互联网推荐、自然语言处理、语音识别、机器人等各种场景。

图7：华为昇腾人工智能生态



数据来源：华为昇腾计算产业白皮书，东吴证券研究所

华为云盘古大模型 3.0 基于鲲鹏和昇腾为基础的 AI 算力云平台，以及异构计算架构 CANN、全场景 AI 框架昇思 MindSpore，AI 开发生产线 ModelArts 等，为客户提供 100 亿参数、380 亿参数、710 亿参数和 1000 亿参数的系列化基础大模型。

盘古大模型致力于深耕行业，打造金融、政务、制造、矿山、气象、铁路等领域行业大模型和能力集，将行业知识 know-how 与大模型能力相结合，重塑千行百业，成为各组织、企业、个人的专家助手。

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/707146052130006056