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

内容目录

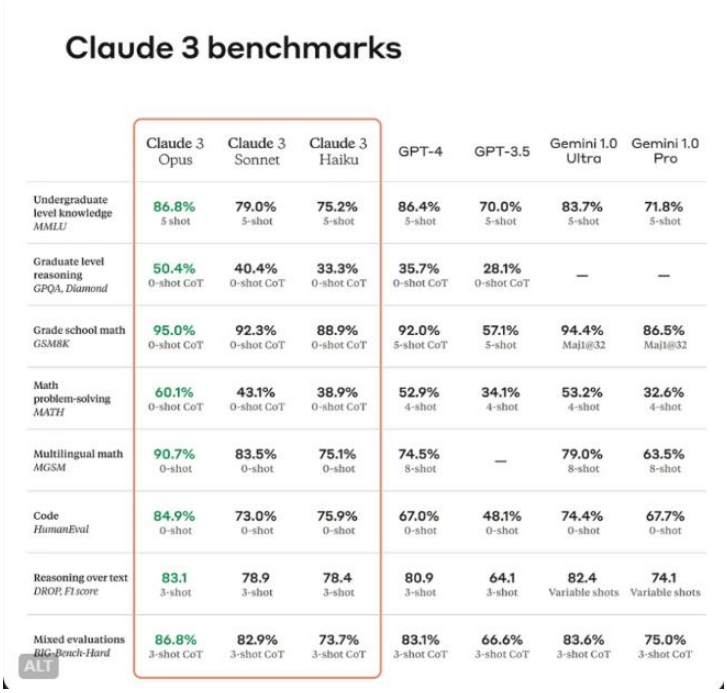
图表目录

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。

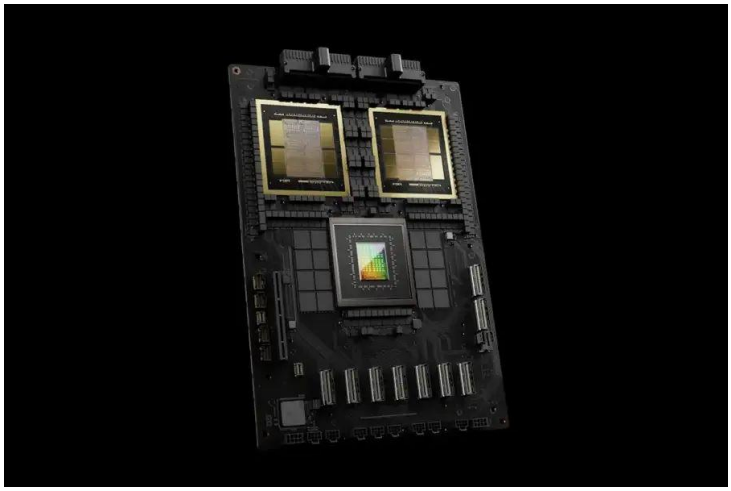
3 月 23 日，媒体援引知情人士透露，OpenAI 计划下周在美国洛杉矶与好莱坞的影视公司和媒体高管会面。OpenAI 希望与好莱坞合作，并鼓励电影制作人将 OpenAI 最新 AI 视频生成工具 Sora 应用到电影制作中，从而拓展 OpenAI 在娱乐行业的影响力。

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。

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。

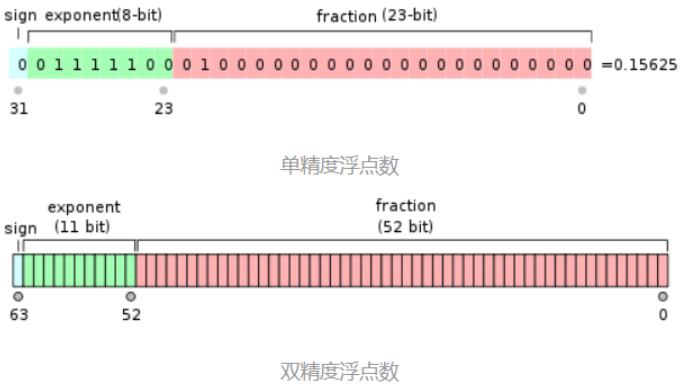
3. 国内算力产业现状盘点

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。

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。

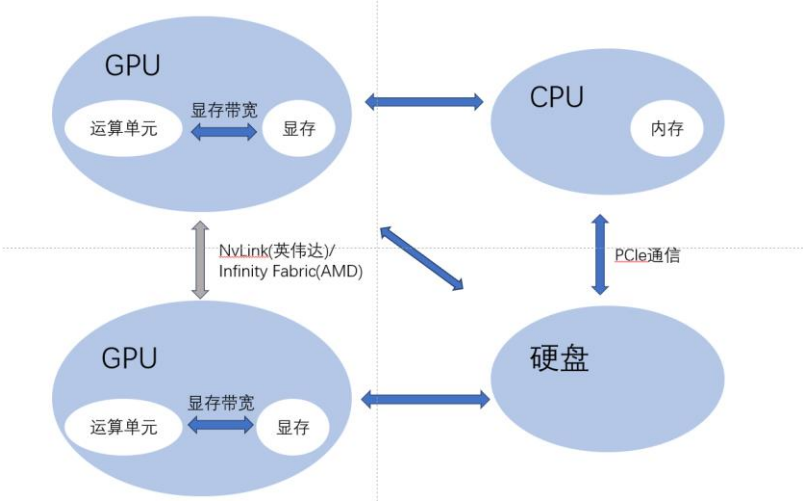
在微调阶段，通常会冻结大部分参数，只训练小部分参数。同时，也会选择非常多的优化技术和较少的高质量数据集来提高微调效果，此时，由于模型已经在预训练阶段进行了大量的训练，微调时的数值误差对模型的影响通常较小。也常常选择 16 位精度训练。因此通常比预训练阶段消耗更低的显存资源。

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

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。

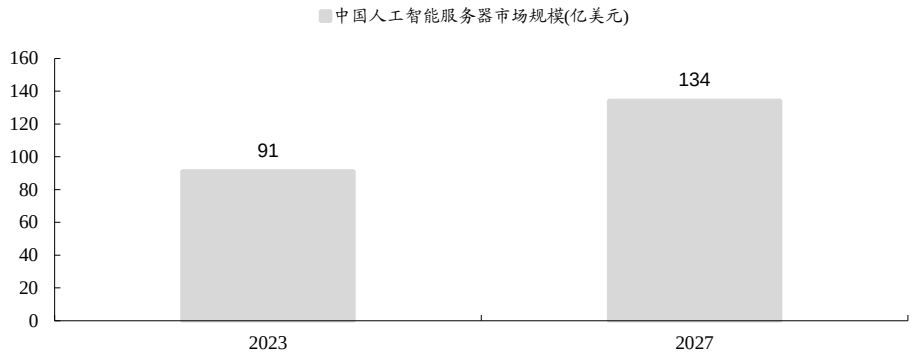
政府关系。国产 AI 芯片的采购一大驱动为政策支持，具有良好的政府关系和客户渠道，可以打开市场空间。

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



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：

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。

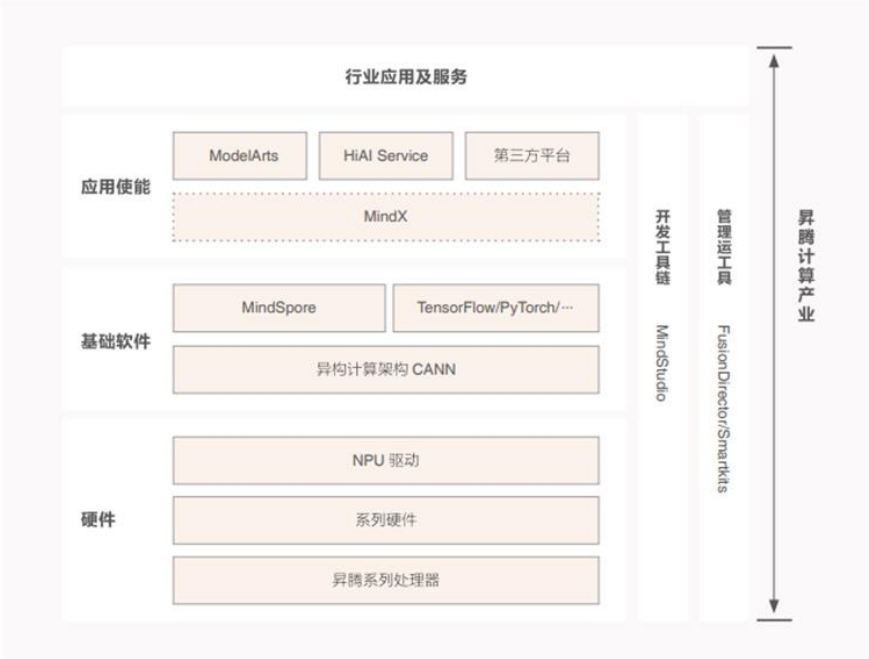
AI 计算框架包括开源的 MindSpore,以及各种业界流行的框架。

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。

华为与行业伙伴一起推动华为大模型行业化。

图8：华为大模型生态合作伙伴

数据来源：华为官方公众号，东吴证券研究所

4.1.1. 昇腾服务器

华为昇腾整机合作伙伴与鲲鹏整机合作伙伴几乎一致，产线共用，从华为直接获取AI服务器或者芯片板卡制造成服务器。

图9：华为昇腾整机合作伙伴主业情况（截至2024年3月24日）

公司名称	主营业务	2022年公司收入和利润	鲲鹏相关业务	当前市值（亿元）
高新发展	建筑施工、功率半导体	收入65.71亿元，归母净利润1.99亿元	拟定增收购华鲲振宇70%股权	299
四川长虹	智能电视、白电、冰箱压缩机等	收入924.82亿元，归母净利润4.68亿元	现穿透后拥有华鲲振宇2.42%股权	268
神州数码	IT分销、云管理、云转售	收入1158.80亿元，归母净利润10.04亿元	100%持有神州鲲泰	216
烽火通信	通信系统设备、光纤及光缆、数据网络产品	收入309.18亿元，归母净利润4.06亿元	持有长江计算80%股权	221
软通动力	IT软件服务外包	收入109.04亿元，归母净利润9.73亿元	100%持有同方股份	490
拓维信息	软件云服务、手机游戏代理发行	收入22.37亿元，归母净利润-10.13亿元	拥有湘江鲲鹏、云上鲲鹏、九霄鲲鹏等控股子公司	207
广电运通	金融科技、智慧城市	收入1158.80亿元，归母净利润10.04亿元	持股16%广电五舟	308

数据来源：公司年报，东吴证券研究所

4.1.2. 昇腾一体机

AI 训推一体机是指将大模型等软件和普通 AI 服务器整合在一起对外销售的整机。

用户画像：主要为 AI 能力自建能力较弱，想要借助 AI 软硬件一体化解决方案构建 AI 能力的客户。

销售方：主要为 ISV，从华为整机厂拿到昇腾整机，然后装上 AI 模型和相关软件直接销售给终端使用客户。

单价：训推一体机由于整合了 AI 大模型等软件产品，单价会明显高于昇腾 AI 服务器裸机，具体价格看软件价格加持价值量。

图10：已发布训推一体机主要产品

合作厂商	名称	算力	简述	时间
华鲲振宇	AI训练开发一体机	2.24 PFLOPS	专为高校和科研院所设计，将AI算力、A平台软件、A开发架开发组件和存储高效融合，构建完整的AI数据服务与开发工作流一体化系统	2023/3/14
软通动力	训推一体化平台	2.5 PFLOPS	基于算聘AI基础硬件平台，整合天鹤OS 操作系统等组件，搭载自有A中台，支持一站式AI开发，为用户提供多种交与式AI模型，深度适配不同A 应用场景	2023/6/5
云从科技	从容大模型训推一体机	2.5 PFLOPS	从容大模型推训一体机结合云从传统视觉优势，可以提供语言、视觉、多模态三大类基础模型推理和训练能力。并基于从容大模型简法及工具，大大降低了用户训练、构建和管理大横型的难度，助力企业打造专属行业大模型，实现 5倍效率提升。	2023/8/2
科大讯飞	星火一体机	2.5 PFLOPS	星火一体机的训练和推理一体化部署，可用于问答系统、对话生成知识图谱构建、智能推荐等多个领域的应用，具备大模型预训练多模态理解与生成、多任务学习和迁移等能力	2023/8/15
智谱AI、华鲲振宇	训推/推理/代码一体机	暂未披露	昇腾基础软硬件平台已与智谱GLM大模型达成深度对接，充分发挥软硬件协同优势，便利开发者和用户，实现普惠AI	2023/9/6
医渡科技	医疗领域专属大模型训推一体机	暂未披露	内置医渡科技全栈自主研发的医疗亚域基础模型，并提供安全高性能计算环境、医渡大模型工具包以及可演示试用场景的API，打造了大模型落地医疗行业的新范式	2023/9/21
安恒信息	大模型一体机	暂未披露	安恒信息恒脑安全垂域大模型已顺利通过A框架异思MindSpore相互兼容性测试认证，基于算膀联合开发的大模型一体机已完成适配	2023/9/22
中软国际	昇腾云+混合云一体机	暂未披露	“异腾云+混合云一体机”是基于馄鹏920系列+异腾910系列的新一代GPU芯片的大模型训练和推理一体机，具备更高算力、极致能效比和高速网络宽带，可用于大模型开发、大模型使用和大模型的运维管理。	2023/10/20

数据来源：各公司公众号，官网，东吴证券研究所

4.2. 海光信息

DCU 已经实现批量出货，迎来第二增长曲线。海光 DCU 以 GPGPU 架构为基础，兼容通用的“类 CUDA”环境，主要应用于计算密集型和人工智能领域。深算二号已经于 Q3 发布，实现了在大数据、人工智能、商业计算等领域的商用，深算二号具有全精度浮点数据和各种常见整型数据计算能力，**性能相对于深算一号性能提升 100%。**

海光 DCU 产品性能可达到国际上同类型主流高端处理器的水平。深算一号采用先进的 7nm FinFET 工艺，能够充分挖掘应用的并发性，发挥其大规模并行计算的能力，快速开发高能效的应用程序。选取公司深算一号和国际领先 GPU 生产商 NVIDIA 公司高端 GPU 产品（型号为 A100）及 AMD 公司高端 GPU 产品（型号为 MI100）进行对比，可以发现典型应用场景下深算一号的性能指标可达到国际同类型高端产品的同期水平。

图11：海光 DCU 深算一号和英伟达 A100 性能对比

项目	海光	NVIDIA
品牌	深算一号	A100
生产工艺	7nm FinFET	7nm FinFET
核心数量	4096（64 Cus）	2560 CUDA processors 640 Tensor processors
内核频率	Up to 1.5GHz（FP64） Up to 1.7Ghz（FP32）	Up to 1.53Ghz
FP 64	10.1 TFLOPS	9.7 TFLOPS
FP 32	11.5 TFLOPS	19.5TFLOPS
FP 16	24.5 TFLOPS	312 TFLOPS
显存容量	32GB HBM2	80GB HBM2e
显存位宽	4096 bit	5120 bit
显存频率	2.0 GHz	3.2GHz
显存带宽	1024 GB/s	2039 GB/s
TDP	350 w	400 w
CPU to GPU互联	PCIe Gen4 x 16	PCIe Gen4 x 16
GPU to GPU互联	xGMIx2, Up to 184 GB/s	NVLink up to 600 GB/s

数据来源：招股说明书，英伟达官网，东吴证券研究所

生态兼容性好。海光 DCU 协处理器全面兼容 AMD 的 ROCm GPU 计算生态，由于 ROCm 和 CUDA 在生态、编程环境等方面具有高度的相似性，CUDA 用户可以以较低代价快速迁移至 ROCm 平台，因此 ROCm 也被称为“类 CUDA”。因此，海光 DCU 协处理器能够较好地适配、适应国际主流商业计算软件和人工智能软件。

海光 DCU 相比海外性价比较高，总体在国内领先。从性能、生态综合来看，海光 DCU 处于国内领先水平，是国产 AI 加速处理器中少数大量销售，且支持全部精度的产品。

在商业应用方面,公司的 DCU 产品已得到百度、阿里等互联网企业的认证,并推出

联合方案,打造全国产软硬件一体全栈 AI 基础设施。

4.3. 寒武纪

寒武纪成立于 2016 年，专注于人工智能芯片产品的研发与技术创新，致力于打造人工智能领域的核心处理器芯片。寒武纪主要产品线包括云端产品线、边缘产品线、IP 授权及软件。

图12：寒武纪主要产品矩阵

智能加速卡	智能加速系统	智能边缘计算模组	终端智能处理器IP	软件开发平台
思元370系列 MLU370-S4/S8智能加速卡 MLU370-X4智能加速卡 MLU370-X8智能加速卡 思元290 MLU290-M5智能加速卡 思元270系列 MLU270-S4智能加速卡 MLU270-F4智能加速卡	玄思1000智能加速器整机	思元220系列 MLU220-SOM智能模组 MLU220-M.2边缘端智能加速卡	Cambricon-1M Cambricon-1H	寒武纪基础软件平台 MagicMind

数据来源：公司官网，东吴证券研究所

寒武纪思元（MLU）系列云端智能加速卡与百川智能旗下的大模型 Baichuan2-53B、Baichuan2-13B、Baichuan2-7B 等已完成全面适配，寒武纪思元（MLU）系列产品性能均达到国际主流产品的水平。

2024 年 1 月 22 日，寒武纪与智象未来 (HiDream.ai) 在北京签订战略合作协议。寒武纪思元（MLU）系列云端智能加速卡与智象未来自研的“智象多模态大模型”已完成适配，在产品性能和图像质量方面均达到了国际主流产品的水平。

4.4. 景嘉微

2024 年 3 月 12 日，公司面向 AI 训练、AI 推理、科学计算等应用领域的景宏系列高性能智算模块及整机产品“景宏系列”研发成功，并将尽快面向市场推广。

景宏系列是公司推出的面向 AI 训练、AI 推理、科学计算等应用领域的高性能智算模块及整机产品，支持 INT8、FP16、FP32、FP64 等混合精度运算，支持全新的多卡互联技术进行算力扩展，适配国内外主流 CPU、操作系统及服务器厂商，能够支持当前主流的计算生态、深度学习框架和算法模型库，大幅缩短用户适配验证周期。

5. 算力租赁

算力租赁就是对算力资源进行出租。使用者可以按需调用算力资源而无需自建算力基础设施。

算力租赁是数字经济时代的新兴产物。算力使用者无需投入大量资金购买计算设备，却可以使用高效稳定的计算服务，并根据实际使用情况支付相应费用。使用者通过租赁计算资源，可以快速地启动项目，减少相应成本。

AI 算力租赁刚刚兴起，参与方众多，格局还比较分散。当前布局 AI 算力租赁市场的主要分为以下几类。1）传统云计算服务提供商，比如三大运营商、阿里、腾讯等；2）具备 IDC 建设运营能力的央国企，比如云赛智联、广电运通等；3）具备 IDC 建设运营相关能力的民企，比如润泽科技、润建股份等；4）跨界厂商，比如迈信林等。

AI 算力租赁目前的核心竞争力是谁能拿到满足客户需求的 AI 算力卡。

国内大模型不断突破，应用不断落地，算力租赁需求有望持续提升。阶跃星辰提到通过自建机房+租用算力，积极进行算力储备。

6. 算力液冷

算力服务器液冷技术是一种采用液体作为散热介质的冷却方式。算力服务器液冷技术主要分为冷板式、浸没式和喷淋式三种。冷板式液冷目前行业成熟度最高，2023 上半年，中国液冷服务器市场中，冷板式占到了 90%。

两大催化推动算力液冷产业加速发展：1） AI 的快速发展，GPU 成为未来数据中心建设的主要方向。GPU 功耗显著高于 CPU，且提升速度逐步加快。3 月 19 日，GTC 大会英伟达提出 GB200 使用液冷方案，其中 GB200 NVL72 服务器提供 36 个 CPU 和 72 个 Blackwell GPU，并使用一体水冷散热方案，全部采用液冷 MGX 封装技术，成本和能耗降低 25 倍。2）国家政策对数据中心 PUE 建设要求越来越高。液冷技术是降低制冷系统能耗的主要技术手段。

液冷技术壁垒不高，行业壁垒较高。算力液冷难点在于修改服务器，服务器往往承载客户核心业务，对稳定性要求较高。服务器厂商对服务器构成和工作情况最为了解，因此服务器厂商具有先天优势。随着市场空间逐步打开，第三方厂商也有望进入市场。

测算：液冷服务器市场空间主要来自于两方面，一方面是存量服务器改造，另一方面是新增服务器建设。

存量改造：

根据《基于价值工程的数据中心液冷与风冷比较分析》数据，浸没式液冷建设成本为 11818 元/kW，我们假设冷板式液冷建设成本约为 4000 元/kw。假设 AI 服务器功耗为 10kW，则对应单台服务器浸没式和冷板式液冷建设成本分别为约为 11 万和 4 万元。

中国电子信息产业发展研究院副院长张小燕介绍，截至 2022 年 Q1，我国在用数据中心机架总规模达到 520 万架，在用数据中心服务器规模达 1900 万台。

假设 2025 年渗透率提升，单价和服务器机架数维持不变。

表1：冷板和浸没式液冷存量改造市场空间测算

	2024E	2025 及以后
渗透率	10%	50%
存量服务器机架（万架）		520
冷板单价（元）		40000
浸没单价（元）		110000
冷板市场空间（亿元）		832
浸没市场空间（亿元）		2288

数据来源：《基于价值工程的数据中心液冷与风冷比较分析》，CEC，电信运营商液冷技术白皮书，东吴证券研究所

新增数量：

前瞻产业研究院预计到 2027 年，中国 AI 服务器出货量将达到 65 万台，2022-2027 年年均复合增长率(CAGR)约为 18%。假设 2027 年全部采用冷板式液冷，则市场规模为 260 亿元。

表2：冷板和浸没式液冷 AI 服务器增量改造市场空间测算

	2027E
中国 AI 服务器出货量（万台）	65
冷板单价（元）	40000
浸没单价（元）	110000
冷板市场空间（亿元）	260
浸没市场空间（亿元）	715

数据来源：前瞻研究院，《基于价值工程的数据中心液冷与风冷比较分析》，东吴证券研究所

7. 全国一体化算力网

2023 年 12 月 29 日，国家发展改革委等部门发布《关于深入实施“东数西算”工程加快构建全国一体化算力网的实施意见》。

国家算力地位提升：引导各类算力向国家枢纽节点集聚，国家枢纽节点外原则上不得新建各类大型及超大型数据中心，坚决避免区域间盲目无序竞争。算力是数字经济的底座，是未来重要的战略资源，国家算力地位提升，算力逐渐基础设施化。并提出到 2025 年底，国家枢纽节点地区各类新增算力占全国新增算力的 60%以上，国家枢纽节点算力资源使用率显著超过全国平均水平。

提升西部算力利用率：积极推动东部人工智能模型训练推理、机器学习、视频渲染、离线分析、存储备份等业务向西部迁移。1ms 时延城市算力网、5ms 时延区域算力网、20ms 时延跨国家枢纽节点算力网在示范区域内初步实现。随着算力时延问题逐步解决，西部算力性价比持续提升。

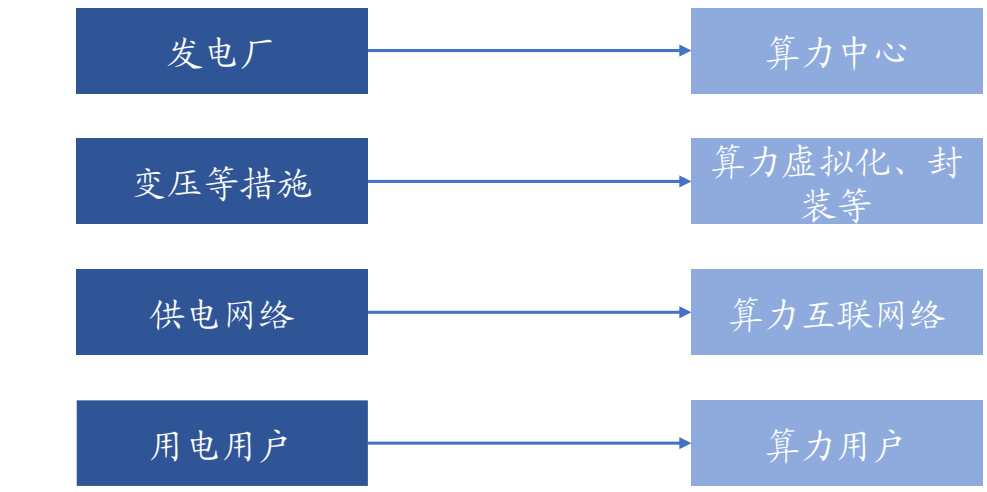
培育算网服务商：支持培育专业化算网运营商，加强算力与网络在运行、管理及维护的全环节运营管理，探索统一度量、统一计费、统一交易、统一结算的标准体系和算网协同运营机制。算力网建设成本投入较大，我们认为算力调度未来会衍生出盈利模式，来覆盖高昂的前期投入成本。政策已经明确提出要培育“专业化算网运营商”，其有望像国家电网一样实现商业模式盈利。

为数据要素市场建设打基础：1）依托国家枢纽节点布局，差异化统筹布局行业特征突出的数据集群，促进行业数据要素有序流通。2）推动各级各类数据流通交易平台利用国家枢纽节点算力资源开展数据流通应用服务，促进数据要素关键信息登记上链、存证备份、追溯溯源。算力是数据要素市场建设的基础。一方面，各行各业数据要素变现、安全等需要大量算力需求；另一方面，算力调度本质是数据的流动，数据要素市场的数据流动可以依托国家算力枢纽。

金融支持：支持产权清晰、运营状况良好的绿色数据中心集群、传输网络、城市算力网、算电协同等项目探索发行REITs。算力网建设需要大量前期资金投入，且在网络建成、规模效应显现之前，企业参与经济性较低。政策明确给予金融支持，有望大幅提升企业参与意愿，加快算力网建设进度。

算力调度类似于电力调度。算力中心可以类比发电厂，算力的用户是大模型、应用等厂商，算力调度就是通过对算力的调度，使得一定范围内算力的需求和供给达到平衡。算力调度网络建成后能够用集群力量弥补国产单芯片能力较弱的短板；提升不同地区的算力利用率；满足客户对异构算力的多样化需求。

图13：算力调度涉及的关键环节



数据来源：中国信通院，东吴证券研究所

算力调度运营是算力调度商业模式最好的环节：算力调度分为基础设施建设、算力调度平台建设、服务运营和算力应用层。算力调度运营是商业模式最好、空间最大的环节。算力调度运营能够坐拥平台流量，帮助客户解决闲置算力，有望实现“抽成”商业模式。从算力调度的范围来看，可以分为全国算力调度、区域算力调度、企业级算力

调度。

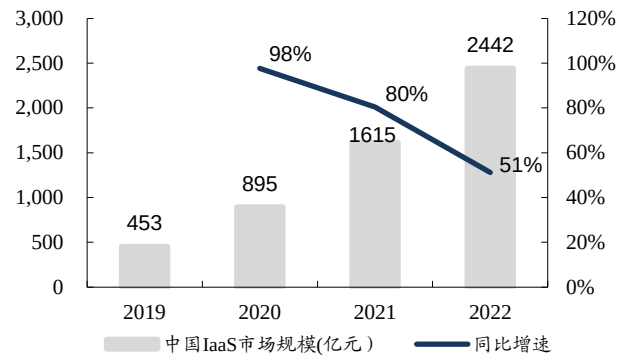
央国企有望在算力调度中大有作为。算力网络运营和调度犹如运营国家电网和石油运输网络，央国企有望承担起算力网络运营和调度的重任。

当前中国算力调度市场处于早期阶段，市场格局较为分散，参与者众多。根据主导方不同，目前主要有四种类型的算力调度平台：运营商主导、政府主导、企业主导、行业机构主导。

算力调度商业模式展望：1）基础设施与算力平台等一次性项目建设和后续运维。当前我国算力网络建设还处于早期阶段，需要算力网络和算力调度平台等基础软硬件建设。2）算力调度平台运营抽成费用。算力调度运营方对接供需双方，将算力利用起来，有望从中抽成。3）算力调度的其他生态费用。算力调度平台不仅仅是提供算力的运营，未来有望进一步发展成为应用商店，客户不但能够购买算力，还能购买相关工具和应用，衍生类似“App Store”的生态费用。

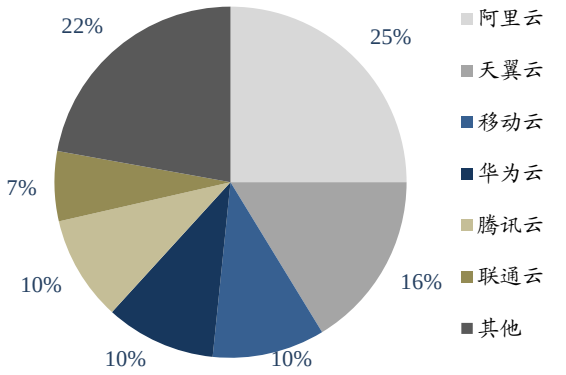
2022 年中国 IaaS 市场规模为 3413 亿元。我们认为 IaaS 市场包括了通用计算、AI 计算和超算。根据中国信通院数据，2022 年中国公有云 IaaS 市场规模为 2442 亿元，我们测算 2022 年中国私有云 IaaS 市场规模为 971 亿元，则 2022 年中国 IaaS 市场规模为 3413 亿元。

图14：2019-2022 年中国 IaaS 市场规模（公有云）



数据来源：中国信通院，东吴证券研究所

图15：2022 年中国公有云 IaaS 市场格局



数据来源：中国信通院，东吴证券研究所

根据《算力基础设施高质量发展行动计划》，中国 2025 年算力规模为 300EFLOPS，是 2023 年的 1.4 倍。假设中国 IaaS 市场规模 2023 年相比 2022 年正增长，则 2025 年中国 IaaS 市场规模为 4778 亿以上。根据中国工程院院士郑纬民、福卡智库等提供数据，2022 年，中国算力利用率仅为 30%，私有云更低。**我们按 2025 年 35%利用率计算，2025 年仅需要盘活的 IaaS 市场空间就有 8873 亿元。**

图16：中国算力基础设施高质量发展指标

	序号	指标	2023 年	2024 年	2025 年
计算力	1	算力规模（EFLOPS）	220	260	300
	2	智能计算中心（个）	30	40	50
	3	智能算力占比（%）	25	30	35

数据来源：工信部等六部门，东吴证券研究所

算力调度运营商能够实现“抽成”收入。算力调度运营商能够连接算力供需，同时提供算力封装和算力计量等服务，我们预计可以从算力租赁费用中收取调度费，实现“抽成”收入。淘宝、京东等电商平台佣金抽成比例一般约为 2%以上，国家电网的输配电价占销售电价约 30%。算力调度比电商平台难度更高，需要提供算力封装等增值服务，抽成比例有望更高，但由于不承担硬件建设成本，应低于输配电价占比。

我们假设悲观、中性和乐观情况下抽成比例分别为 5%、8%、10%。我们测算对应 2025 年中国算力调度潜在市场规模为 444、710、887 亿元。

表3：2025 年中国算力调度潜在市场规模测算

2022 年中国 IaaS 市场规模(亿元)	3413		
2025 年中国 IaaS 市场规模（亿元）	4778		
2025 年中国算力利用率	35%		
2025 年中国需盘活算力规模(亿元)	8873		
调度费率	5%	8%	10%
2025 年潜在调度费用市场规模(亿元)	444	710	887

数据来源：中国信通院，国家电网，中国工程院，东吴证券研究所

8. 央企 AI

2 月 19 日，国务院国资委召开“AI 赋能 产业焕新”中央企业人工智能专题推进会。国务院国资委党委书记、主任张玉卓在会上讲话强调，要深入学习贯彻习近平总书记关于发展人工智能的重要指示精神，推动中央企业在人工智能领域实现更好发展、发挥更大作用。

会议认为，加快推动人工智能发展，是国资央企发挥功能使命，抢抓战略机遇，培育新质生产力，推进高质量发展的必然要求。

会议强调，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。要夯实发展基础底座，把主要资源集中投入到最需要、最有优势的领域，加快建设一批智能算力中心，进一步深化开放合作，更好发挥跨央企协同创新平台作用。开展 AI+专项行动，强化需求牵引，加快重点行业赋能，构建一批产业多模态优质数据集，打造从基础设施、算法工具、智能平台到解决方案的大模型赋能产业生态。

9. 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。

10. 风险提示

政策支持不及预期。公共算力平台建设和运营，以及 AI 芯片国产化需要政策强力支持，如果政策支持力度不及预期，行业发展进度会受影响。

技术研发不及预期。AI 芯片技术壁垒较高，国产化难度较大，如果国产 AI 芯片研发进度不及预期，将会影响国产化进度。

AI 发展不及预期。AI 算力需求增加主要系 AI 技术快速发展所致，如果 AI 技术发展减缓，则对算力的需求会有所降低。

东吴证券 SOOCHOW SECURITIES

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券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并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn