

智能算力产业发展 白皮书

2024年7月

发布单位

国家信息中心大数据发展部、上海人工智能研究院、东方证券-上海人工智能研究院联合实验室

特别鸣谢

广州数据集团、广电运通、粤港澳大湾区大数据研究院

编制组成员

专家指导组成员

于施洋 国家信息中心大数据发展部主任

宋海涛 上海人工智能研究院院长

陈刚 东方证券首席研究总监

编制组成员

国家信息中心大数据发展部

郭明军 规划与应用处处长

童楠楠 规划与应用处副处长

陈东 规划与应用处高级工程师

窦悦 规划与应用处副研究员

郭巧敏 规划与应用处助理研究员

赵正 规划与应用处助理研究员

马文博 规划与应用处助理研究员

马骁 规划与应用处实习研究员

张琳颖 规划与应用处实习研究员

上海人工智能研究院

杨浩 副院长

孙丽 数字经济研究中心主任

苏合检 算力算网中心副主任

唐晟凌 数字经济研究中心资深咨询师

康起明 数字经济研究中心资深咨询师

徐艺 数字经济研究中心资深咨询师

仲惠颖 数字经济研究中心资深咨询师

东方证券-上海人工智能研究院联合实验室

浦俊懿 计算机行业首席分析师，东方证券-上海人工智能研究院联合实验室主任

宋鑫宇 计算机行业分析师

陈超 计算机行业前瞻科技首席分析师

杜云飞 计算机行业高级分析师

覃俊宁 计算机行业分析师

粤港澳大湾区大数据研究院

李琳 常务副院长

王天瑞 工程咨询中心高级咨询师

栾雪洋 工程咨询中心高级咨询师

目录



一、AI大模型时代到来，智能计算成为关键引擎	04
1.1 AI大模型取得突破，算力需求正从通用计算转向智能计算	04
1.2 智能算力需求将呈现指数级增长	08
1.3 智算产业上升为国家战略竞争点，国家与地方政策加大智算产业支持力度	11
1.3.1 国家层面制定一系列政策，在全国范围推动智能算力发展	12
1.3.2 各地积极布局算力，智算中心建设热度持续	15
二、智算产业蓬勃发展，产业生态逐步形成	23
2.1 智算中心建设迅速落地，多方主体积极参与	23
2.2 智算产业链正在形成	25
三、智算芯片是智算产业核心环节，也是全球科技巨头竞逐热点	27
3.1 智算芯片是中美科技博弈重要领域，国产智算芯片发展刻不容缓	27
3.1.1 智算芯片是智算产业的核心环节	27
3.1.2 智算芯片成为中美科技博弈的焦点之一	28
3.1.3 国产智算芯片发展刻不容缓	31
3.2 海外：英伟达目前独大，AMD、微软、特斯拉、谷歌等开始发力	33
3.2.1 英伟达目前在智算芯片领域独占鳌头	33
3.2.2 AMD在智算芯片领域不断发力	34
3.2.3 微软自研芯片垂直整合Azure软硬件	35
3.2.4 特斯拉推出FSD芯片以最大化其自动驾驶技术潜力	36
3.2.5 谷歌TPU持续迭代性能升级	37
3.3 国内：华为、海光信息等科技大厂积极布局，寒武纪、天数智芯等创业型企业纷纷兴起	38
3.3.1 华为打造全栈自主AI基础软硬件	38
3.3.2 海光DCU依托开放生态加速推广	42
3.3.3 寒武纪云边端协同发展	43
3.3.4 天数智芯立足高性能通用GPU加速构建AI自主生态	44
3.3.5 瀚博半导体芯片产品聚焦AI加速和视频处理	45
3.4 国内厂商在产品性能差距上正在缩小，但应用生态差距较大	48
3.4.1 国产智算芯片在性能方面与英伟达的通用GPU的差距正在缩小	48
3.4.2 国产智算芯片在应用生态方面还有很大发展空间	49
四、智算产业发展趋势	50
4.1 多元异构算力融合与国产化是方向与趋势	50
4.2 智算建设与东数西算工程将有效融合	53
4.3 通过算力调度，智算由单个节点向多点融合与协同演变	55
4.4 智算中心低碳化发展需要算力-电力协同	60

五、优秀案例介绍.....63

5.1 中科曙光.....63

5.2 广州数字科技集团.....66

六、智算发展展望与政策建议69

6.1 加强智算资源整合与布局优化.....69

6.2 加快培育国产软硬件生态.....69

6.3 促进产业链协同与创新升级.....69

6.4 加强智算中心运营保障 70



一、AI大模型时代到来，智能计算成为关键引擎

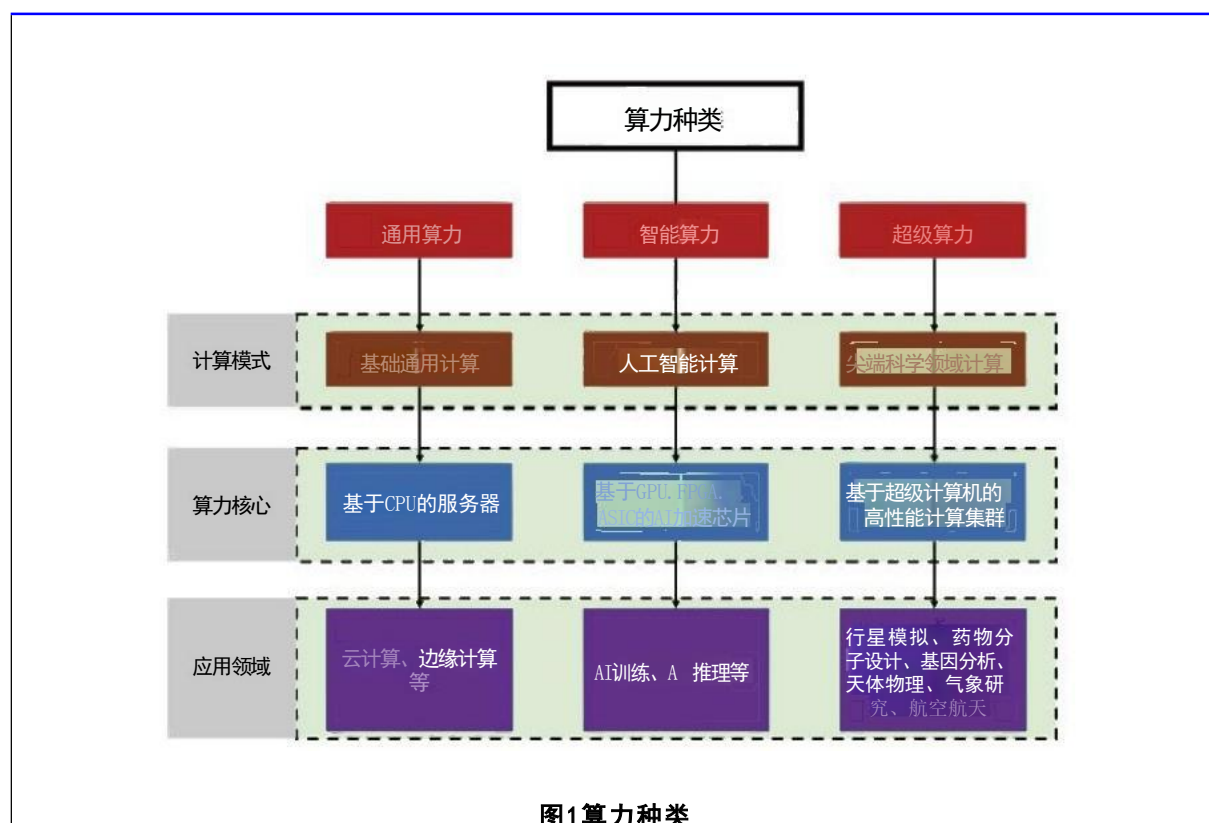
1.1 AI大模型取得突破，算力需求正从通用计算转向智能计算

智能计算是指面向人工智能应用(包括训练、推理、衍生应用场景等)的高性能计算。算力可分为：通用算力、智能算力和超级算力，分别对应三种计算模式：基础计算、智能计算和超级计算。不同应用场景下所需的计算精度不同，通常会采用不同种类的算力。目前，随着人工智能的快速发展以及国家层面的政策助推，智能算力规模和占比越来越大，算力需求逐步从通用计算转向智能计算。

①**通用算力**：基于CPU 芯片的服务器所提供的算力，主要用于计算复杂度适中的云计算、边缘计算类场景，通常这些场景对实时性有一定要求，不适合完全将本地数据搬到异地计算，如移动计算和物联网等。

②**智能算力**：基于GPU、FPGA、ASIC等AI芯片的加速计算平台提供的算力，主要用于人工智能的训练和推理计算，智算中心可以根据不同细分领域业务的算力需求匹配相应的计算能力。对于人工智能的模型训练及推理来说，处理文字、语音、图片或视频等需求较大，单精度、半精度、甚至整型的智能计算能够满足应用需要。

③**超级算力**：由超级计算机等高性能计算集群所提供的算力，主要用于尖端科学领域的计算，比如行星模拟、药物分子设计、基因分析、天体物理、气象研究、航空航天等需要复杂运算、高性能双精度算力的高精尖科研领域。同时，不同超级计算机的处理器、加速卡、框架等各不相同，商业化服务门槛高。



数据来源：公开资料整理



数据来源：华为

驱动力1:通用大模型。机器学习进入大模型时代，ChatGPT 等通用大模型的训练迭代极大拉动对智能算力的需求，模型成功部署后的推理阶段也将需要大量智能算力做支撑。从模型训练角度来说，据J. Sevilla等发布的文章《Compute Trends Across Three Eras of Machine Learning, "2022 International Joint Conference on Neural Networks(IJCNN)》，机器学习的训练计算大概可以分为三个时期：1)第一个时期为2012年之前，训练算力大致遵循摩尔定律，约每20个月翻一番；2)进入深度学习时代，算力翻倍的速度加速至5-6个月；3)2015-2016年左右开启了大模型时代，在这个时期，计算量增长变慢，翻倍时间约为10个月，但整体的训练计算量比深度学习时代的系统大2到3个数量级(OOM)。从2022年底，随着ChatGPT的发布，新一代AI浪潮来临，Bert、GPT4、文心一言等国内外通用大模型相继发布。这些大模型的训练需要千亿、甚至万亿级参数，以及上千GB的高质量数据，大模型的训练迭代极大地拉动了智能算力的需求。另外，随着模型的成熟落地和推广，模型推理所需的智能算力也将逐渐增加，占比不断提高。

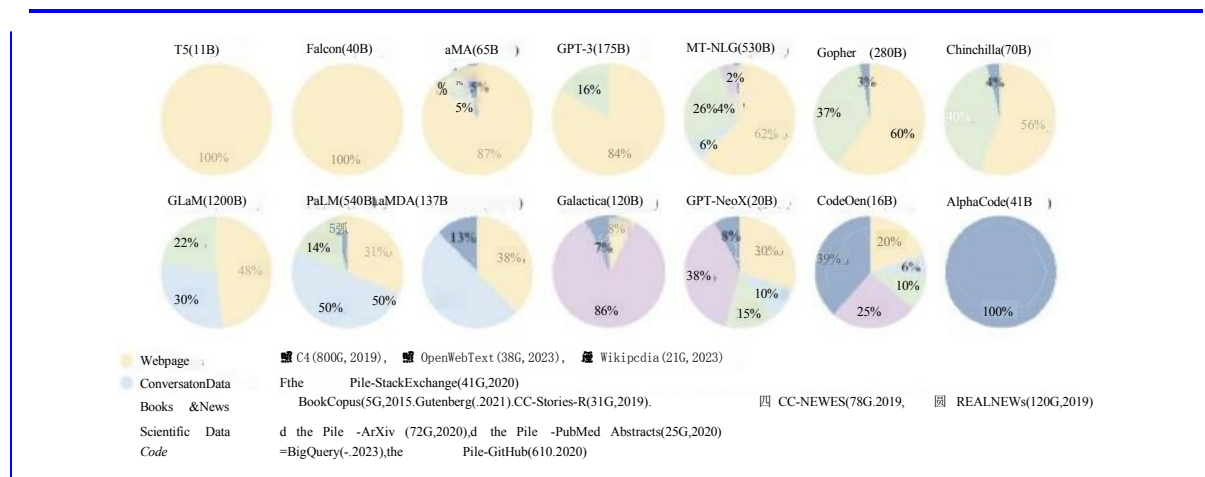
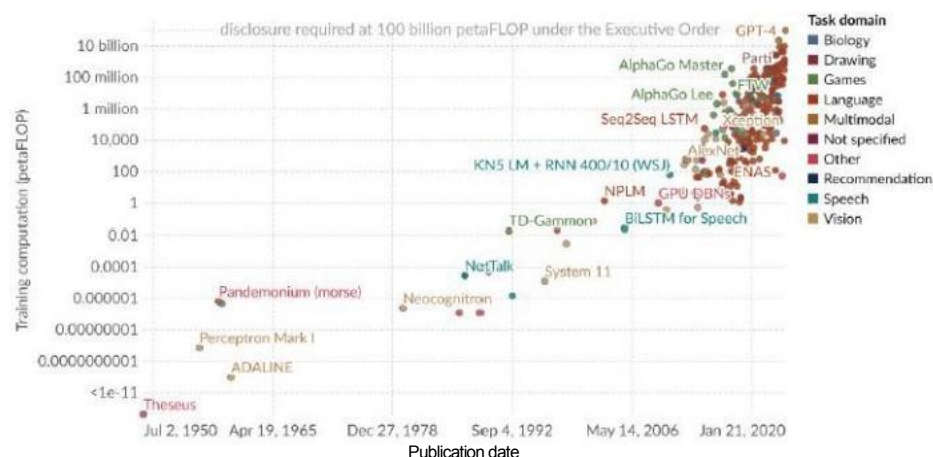


图3大模型参数量及其训练所需高质量数据源构成

数据来源: 论文《A Survey of Large Language Models》

Computation used to train notable artificial intelligence systems

Computation is measured in total petaFLOP, which is 10^{15} floating-point operations estimated from AI literature, albeit with some uncertainty. Estimates are expected to be accurate within a factor of 2, or a factor of 5 for recent undisclosed models like GPT-4.



Data source: Epoch(2023)

OurWorldInData.org/artificial-intelligence(CC BY)

Note: The Executive Order on AI refers to a directive issued by President Biden on October 30, 2023, aimed at establishing guidelines and standards for the responsible development and use of artificial intelligence within the United States.

1 Floating-point operation: A floating-point operation (FLOP) is a type of computer operation. One FLOP is equivalent to one addition, subtraction, multiplication, or division of two decimal numbers.

图4用于训练人工智能系统的算力需求

数据来源: Our World in Data

驱动力2: 垂直行业大模型。垂直行业大模型的训练和基于通用大模型微调的行业应用都需要大量的智能算力做支撑。垂直行业的大模型训练需要大量的智能算力, 另外, 基于大模型的多场景应用也不断拓展。AI渗透千行百业, 拉动智能算力规模高速增长。2022年, 各行各业的AI应用渗透度都呈不断加深的态势, 尤其是在金融、电信、制造以及医疗领域, 为实现业务增长、保持强大竞争力、从而占据更大的市场份额, 企业纷纷入局AI领域, 通过新技术提升传统业务用户体验, 人工智能应用增长迅速。据IDC和浪潮信息联合发布的《2022-2023中国人工智能计算力发展评估报告》, 预计到2023年年底, 中国将有50%的制造业供应链环节采用人工智能技术实现业务体验提升。在未来, 随着AI技术对传统行业赋能作用日益凸显, 更大规模的智算需求出现成为必然。



图5中国人工智能行业应用渗透度 (%) 及同比增长

数据来源: IDC, 浪潮信息

驱动力3:海量数据。随着云计算、大数据、AI、物联网等技术应用发展,数据量呈指数级增长,智能算力需求也随之快速增长。据IDC 最新发布Global DataSphere 2023数据,2023年我国数据量规模预计为30.0ZB,到2027年数据量规模则将达到76.6ZB,2023-2027 年CAGR 达到26.4%。英特尔的研究也表明,AI模型的计算量每年将会增长10倍,这一趋势随着未来人工智能与传统产业的深度融合会进一步加快。自2012年开始,驱动AI的底层机器学习技术计算量呈指数级增长,远超摩尔定律带来的算力提升速度,对以CPU 为主的通用算力提出严峻考验。目前CPU 的物理工艺、核心数已接近极限,但数据量的增长却不会停止。面对激增的数据量,算力不足的问题日益突出。因此,在AI时代下,仅由CPU 作为算力的提供者已经不能满足需求,向包括GPU、FPGA、ASIC 等加速芯片异构而成的智能算力演化成为趋势,并最终将会成为数字经济时代下算力的主角。

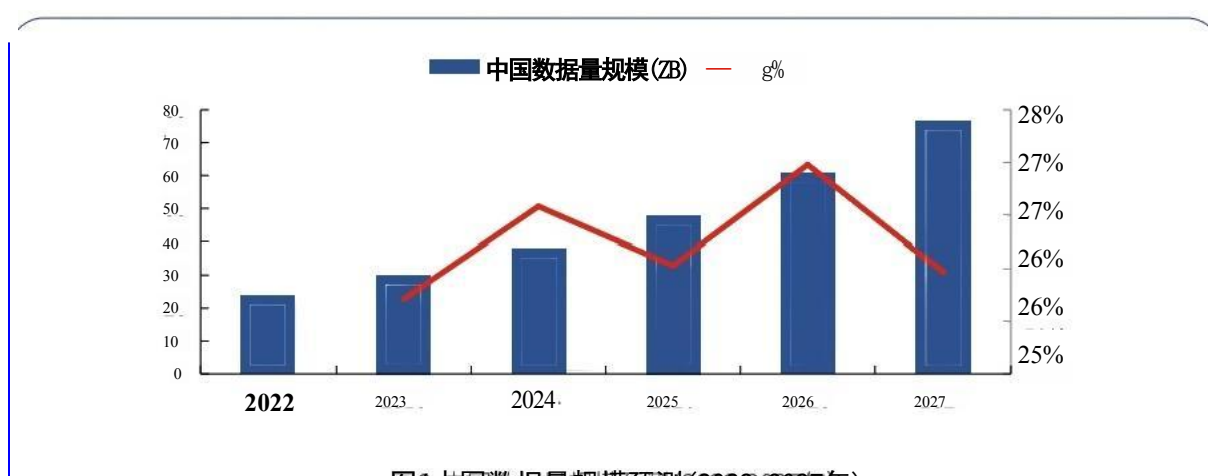


图6中国数据量规模预测(2022~2027年)

数据来源: IDC

当前,我国智能算力占比已经超过通用算力,成为整体算力增长的主要驱动力。我国智能算力占总体算力的比重增长迅速,据中国信通院《中国算力发展指数白皮书(2023)》数据,2017-2022年我国算力内部结构中智能算力占比持续增加,2022年我国算力结构为通用算力:智能算力:超级算力=40%:59%:1%,智能算力占比已较大幅度超过通用算力占比,逐步成为总体算力增长的主要驱动力。据中国信通院预计,未来五年,全球算力规模增速超过50%。

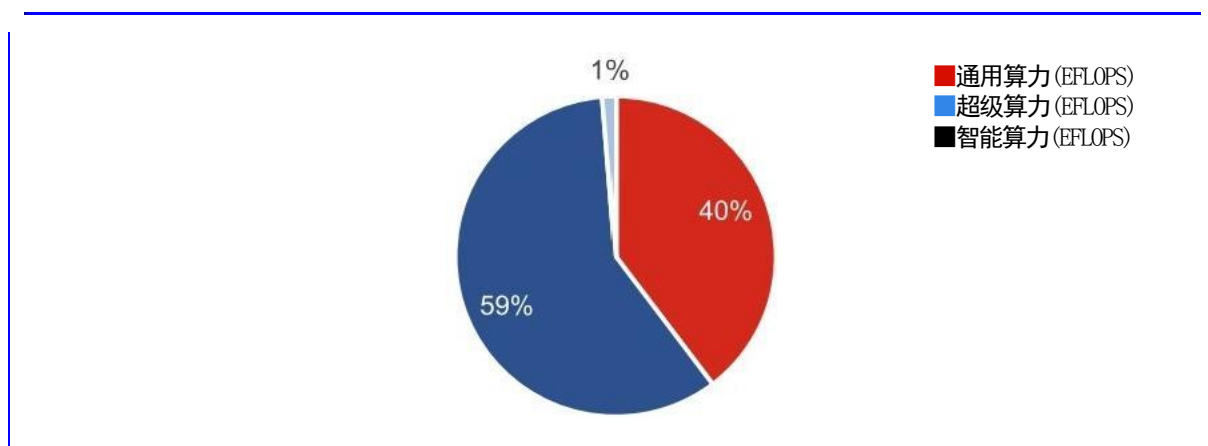


图7 2022年我国算力结构

数据来源: 中国信通院

1.2 智能算力需求将呈现指数级增长

全球范围内算力结构向智能算力集中，预计未来智能算力需求激增。随着数字时代的来临，计算场景愈加丰富，人工智能、物联网等技术飞速发展，数据高速增长，全球范围内智能算力需求激增。据中国信通院数据，2022年全球计算设备算力总规模达到906EFLOPS,同比增速47%，算力结构为通用算力：智能算力：超级算力=48%:50%:2%，智能算力占据一半份额。

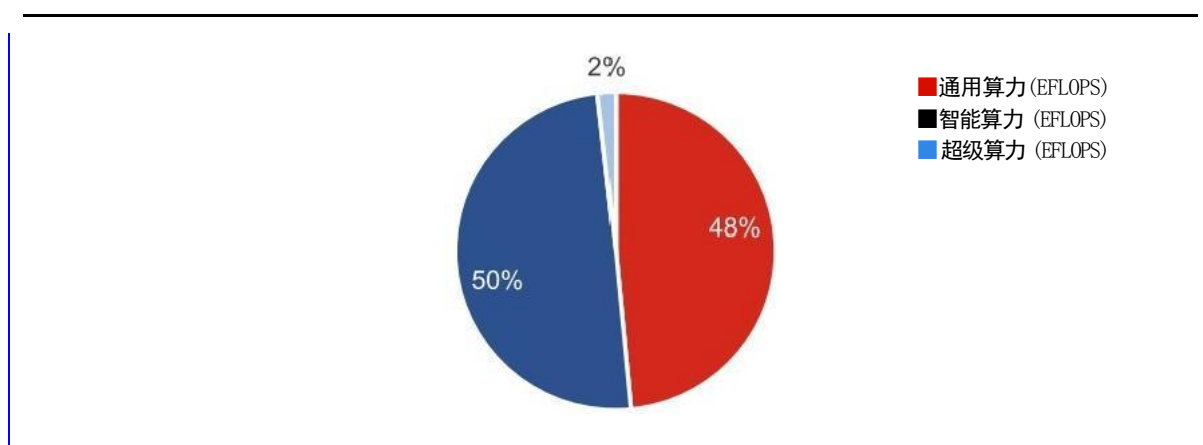


图8 2022年全球算力结构

数据来源：中国信通院

我国智能算力规模高速增长，预计2026年智能算力规模突破ZFLOPS 量级，增长速度远超通用基础算力。在大模型取得突破、应用场景广泛开拓与深入发展的背景下，智能算力需求将在未来几年迎来井喷。随着数据量高速增长、数据要素化进程推进、同时算力模型复杂度日益提升，智能算力作为释放数据价值的必要工具，其需求与规模快速增长。OpenAI 分析显示，从2012年以来，最大规模的AI模型训练中所需要的计算量，每3.5个月便翻倍一次。相比于摩尔定律2年的倍增期，算力需求具有远高于芯片承载计算量的增长速度。

●我国智能算力规模增速预计将远超通用算力规模增速，智能算力呈现井喷式增长态势。据IDC 数据，2020年中国智能算力规模为75.0EFLOPS，之后则保持快速增长态势，预计2023年智能算力规模将为414.4EFLOPS，到2027年预计则将达到1117.4EFLOPS，实现量级突破，进入ZFLOPS 级别，2020-2027年预计CAGR 达到47.1%。与之对比，通用算力规模增速则较慢，预计2023和2027年我国通用算力规模分别为59.3 EFLOPS 和117.3 EFLOPS, 2020-2027 年CAGR 预计为16.8%。



图9 2020-2027年中国智能算力规模及预测 (EFLOPS, 基于FP64 计算)

数据来源: IDC



图10 2020-2027年中国通用算力规模及预测 (EFLOPS, 基于FP16 计算)

数据来源: IDC

●从国家政策规划层面来看，2023年10月工信部等六部委发布《算力基础设施高质量发展行动计划》目标规划到2025年，算力规模超过300EFLOPS,智能算力占比达到35%, 达到105EFLOPS。

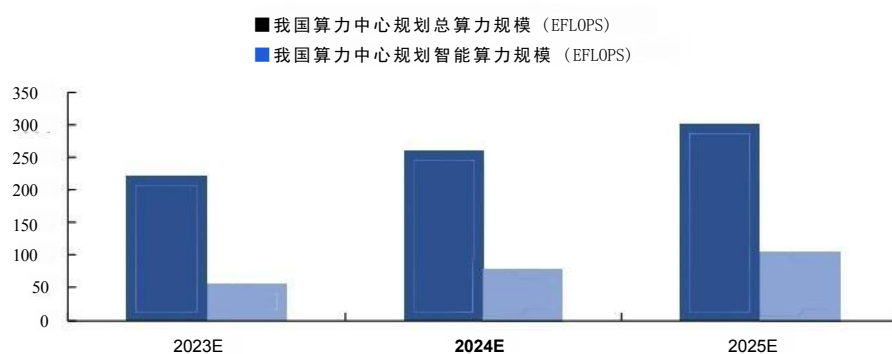


图11 我国规划智算中心算力需求规模

数据来源: 国务院官网

从智能算力总额来看，美、中处于领先地位，而从人均智能算力的角度看，中国仍处于全球中等水平。据《中国算力指数发展白皮书(2022)》，美、中的智能算力处于全球领先地位，分别占全球比重的45%和28%。然而从人均算力的高低来衡量，美国、英国、德国等国家的人均算力普遍高于1000GFlops，而我国的人均算力处于中等水平。据IMD 研究发现，人均算力的水平与一国的智能化水平高度相关，我国积极发展智能算力、打造智算中心是打造国际竞争力、发展综合国力的关键。



图10 2020-2027年中国通用算力规模及预测 (EFLOPS, 基于FP64计算)

数据来源：IDC

●从国家政策规划层面来看，2023年10月工信部等六部委发布《算力基础设施高质量发展行动计划》，目标规划到2025年，算力规模超过300EFLOPS,智能算力占比达到35%, 达到105EFLOPS。

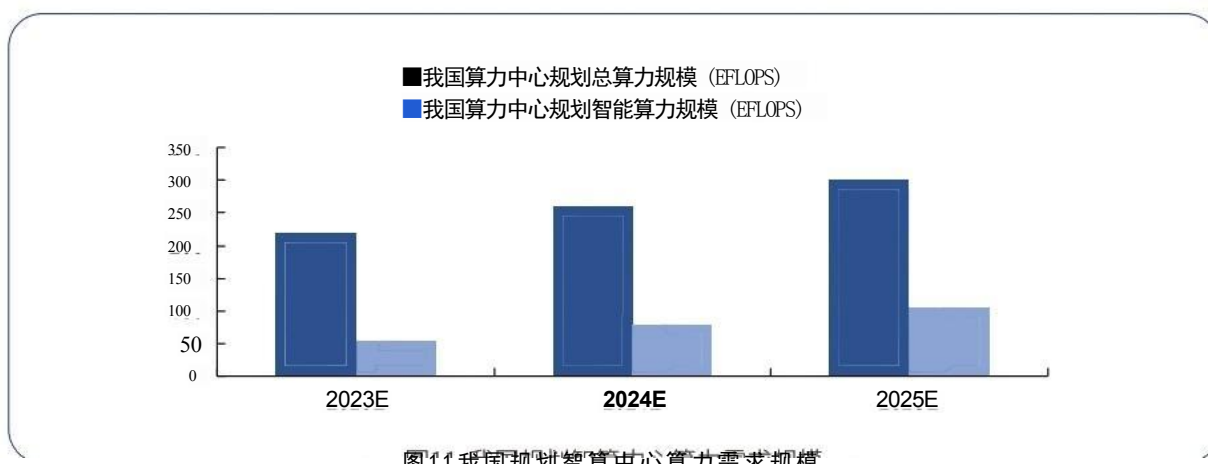


图11 我国规划智算中心算力需求规模

数据来源：国务院官网

从智能算力总额来看，美、中处于领先地位，而从人均智能算力的角度看，中国仍处于全球中等水平。据《中国算力指数发展白皮书(2022)》，美、中的智能算力处于全球领先地位，分别占全球比重的45%和28%。然而从人均算力的高低来衡量，美国、英国、德国等国家的人均算力普遍高于1000GFlops，而我国的人均算力处于中等水平。据IMD 研究发现，人均算力的水平与一国的智能化水平高度相关，我国积极发展智能算力、打造智算中心是打造国际竞争力、发展综合国力的关键。



→图12世界主要国家人均算力水平

数据来源：罗兰贝格

1.3 智算产业上升为国家战略竞争点，国家与地方政策加大智算产业支持力度

智算中心是“东数西算”的关键，在未来将会为社会提供主要智算资源。据国家信息中心，未来80%的场景都将基于人工智能，所占据的算力资源将主要由智算中心提供。AI大模型已经成为国家、企业和科研院所积极发展、重点投入的大方向，而普惠大众的智能算力就是AI发展的基础资源，将要像水、电一样驱动科技发展、社会进步，智算中心则是实现这些科技创新的源泉。人工智能一直是中美两国科技竞争的重要领域，而中美在AI的三大技术数据、算法、算力中各有优势。中国的数据优势较为明显，但是算法和智能算力却明显落后于美国。发展智能算力对我国发展AI与数据要素，推动社会全方位发展有重大战略意义。

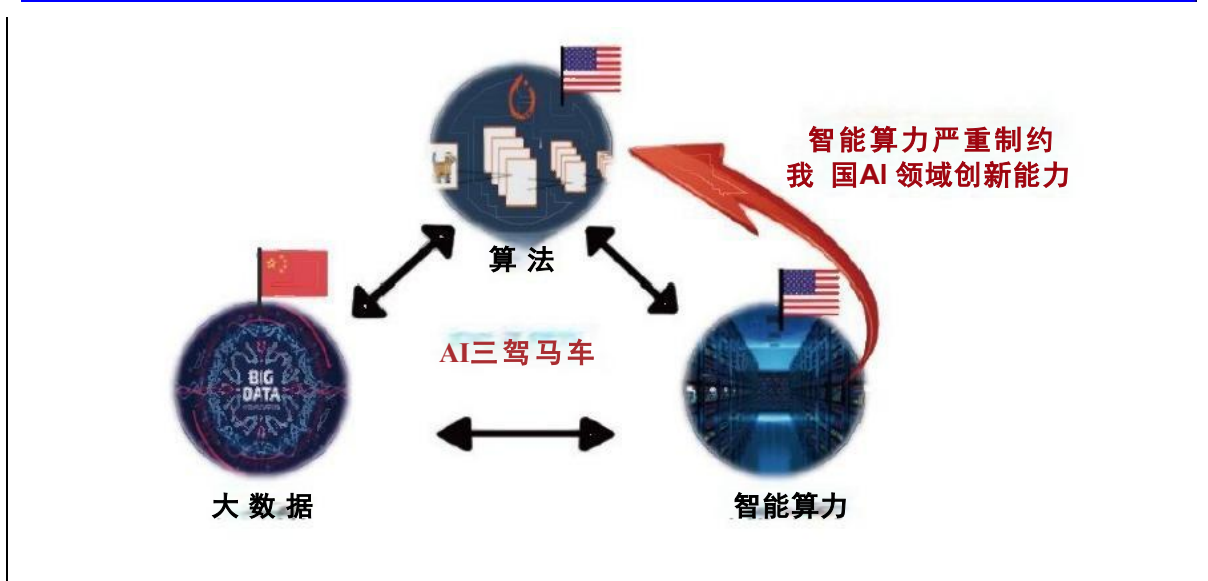


图13 AI的三驾马车——数据、算法、智能算力

资料来源：郑纬民院士学术报告《人工智能算力基础设施的设计、评测与优化》

1.3.1 国家层面制定一系列政策，在全国范围推动智能算力发展

近年来，国家制定一系列政策，在全国范围内推动信息基础设施建设，智算中心的建设是其中的关键环节。

①2020年4月20日，国家发展改革委首次明确新型基础设施范围，将智能计算中心作为算力基础设施的重要代表纳入信息基础设施范畴。2021年5月，国家发改委等四部门联合发布了《全国一体化大数据中心协同创新体系算力枢纽实施方案》，提出布局全国算力网络枢纽节点。

②2022年2月17日，国家发改委、中央网信办、工业和信息化部、国家能源局联合印发通知，同意在京津冀、长三角、粤港澳大湾区、成渝、内蒙古、贵州、甘肃、宁夏等8地启动建设国家算力枢纽节点，并规划了10个国家数据中心集群。“东数西算”工程正式全面启动。而智算中心承载的以模型训练为代表的非实时性计算尤为适合“东数西算”工程。

③科技部出台政策推动公共算力平台建设。依照《新一代人工智能发展规划》，科技部启动“人工智能驱动的科学研究”(AI for Science)专项部署工作，同时发布《科技部办公厅关于开展国家新一代人工智能公共算力开放创新平台申报工作的通知》，提出要推进AI领域的模型与算法创新工作，加快推动国家新一代人工智能公共算力开放创新平台建设，支持高性能计算中心与智算中心异构融合发展。

布局图



图14 “东数西算”八大算力枢纽和十大数据集群布局

资料来源：国际科技创新中心

表1智能算力相关国家/地方政策

发布时间	颁布部门	政策	主要内容
2023年12月	国家发改委等五部门	《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》	统筹通用算力、智能算力、超级算力协同计算。到2025年底，综合算力基础设施体系初步成型。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；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。
2023年10月	工信部等六部门	《算力基础设施高质量发展行动计划》	到2025年，计算力方面，算力规模超过300EFLOPS, 智能算力占比达到35%。
2023年2月	国务院	《数字中国建设整体布局规划》	优化算力布局，促进东西部算力联动，引导通用数据中心、超算中心、智能计算中心、边缘数据中心等合理梯次布局。
2022年2月	国家发改委等部门	《关于同意京津冀地区启动建设全国一体化算力网络国家枢纽节点的复函》	同意在京津冀地区启动建设全国一体化算力网络国家枢纽节点，发展高密度、高效率、低碳数据中心集群。积极承接北京等地实时性算力需求，引导温冷业务向西部迁移。
2022年1月	国务院	《“十四五”数字经济发展规划》	加快构建算力、算法、数据、应用资源协同的全国一体化大数据中心体系，建设数据中心集群。
2021年12月	中央网络安全和信息化委员会	《“十四五”国家信息化规划》	加快构建全国一体化大数据中心协同创新体系，建设京津冀、长三角、粤港澳大湾区、成渝等全国一体化算力网络国家枢纽节点。统筹建设面向区块链和人工智能等的算力和算法中心，构建具备周边环境感应能力和反馈回应能力的边缘计算节点，提供低时延、高可靠、强安全边缘计算服务。
2021年11月	工信部	《“十四五”大数据产业发展规划》	加快构建全国一体化大数据中心体系，推进国家工业互联网大数据中心建设，强化算力统筹智能调度，建设若干国家枢纽节点和大数据中心集群。建设高性能计算集群，合理部署超级计算中心。

表1智能算力相关国家/地方政策

发布时间	颁布部门	政策	主要内容
2021年7月	国家发改委等五部门	《新型数据中心发展三年行动计划(2021-2023年)》	用3年时间，基本形成布局合理、技术先进、绿色低碳、算力规模与数字经济增长相适应的新型数据中心发展格局。技术能力明显提升，产业链不断完善，国际竞争力稳步增强。
2021年5月	国家发改委等四部门	《全国一体化大数据中心协同创新体系算力枢纽实施方案》	提出布局全国算力网络枢纽节点。
2021年3月	全国人大	《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》	加快构建全国一体化大数据中心体系，强化算力统筹智能调度，建设若干国家枢纽节点和大数据中心集群，建设E级和10E级超级计算中心。
2021年1月	工信部	《工业互联网创新发展行动计划(2021-2023年)》	推动工业互联网大数据中心建设，打造工业互联网大数据中心综合服务能力，到2023年基本建成国家工业互联网大数据中心体系，建设20个区域级分中心和10个行业级分中心。
2023年7月	河南	《河南省重大新型基础设施建设提速行动方案(2023—2025年)》	加快建设郑州、洛阳等全栈国产化智能计算中心，构建中原智能算力网。持续提升国家超算郑州中心超算能力，建设智算中心和郑州城市算力网调度中心，综合算力性能保持国际前列，资源利用率达到70%。到2025年智算和超算算力规模超过2000PFLOPS(每秒浮点运算次数)，高性能算力占比超过30%
2023年6月	上海	《临港新片区加快构建算力产业生态行动方案》	目标到2025年，临港新片区将形成以智算算力为主、基础算力和超算算力协同的多元算力供给体系，总算力超过5EFLOPS(FP32)，AI算力占比达到80%，算力产业总体规模突破100亿元。
2023年5月	深圳	《深圳市加快推动人工智能高质量发展高水平应用行动方案(2023-2024年)》	坚持供给智能算力一大要素，强化智能算力集群供给，打造多层次智能算力集群。打造大湾区智能算力枢纽建设企业级智能算力平台。

表1智能算力相关国家/地方政策

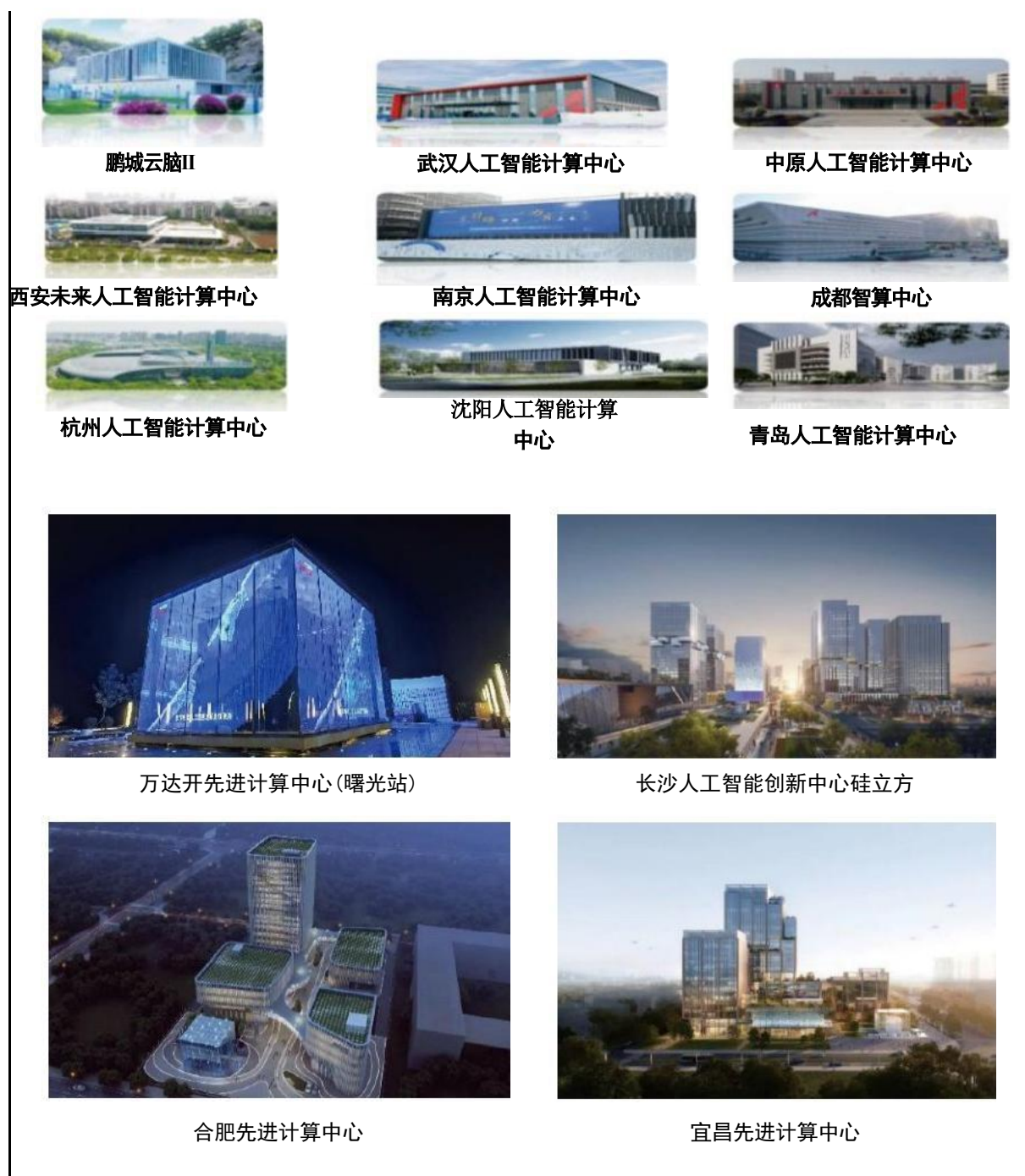
发布时间	颁布部门	政策	主要内容
2023年5月	北京	《北京市促进通用人工智能创新发展的若干措施》	在算力资源统筹供给方面，提出加强与头部公有云厂商等市场主体合作，实施“算力伙伴计划”，并加快推动北京人工智能公共算力平台、北京数字经济算力中心等项目的建设，支撑千亿级参数量的人工智能模型研发，建设统一的多云算力调度平台，方便企业在不同云环境上无缝、经济、高效地运行各类人工智能计算任务。
2023年5月	北京	《北京市加快建设具有全球影响力的人工智能创新策源地实施方案（2023-2025年）》	人工智能算力布局初步形成，国产人工智能芯片和深度学习框架等基础软硬件产品市场占比显著提升，算力芯片等基本实现自主可控。国产硬件比例显著提高，全面兼容国产深度学习框架。人工智能算力资源并网互联，推动基础软硬件实现高质量自主可控。
2023年1月	四川	成都市围绕超算智算加快算力产业发展的政策措施	建立以“算力券”为核心的算力中心运营统筹结算分担机制，结合区块链等新技术实现“算力券”有效监管。每年发放总额不超过1000万元的“算力券”，用于支持算力中介服务机构、科技型中小微企业和创客、科研机构、高校等使用国家超算成都中心、成都智算中心算力资源。

数据来源：政府各部门官网整理

1.3.2各地积极布局算力，智算中心建设热度持续

随着全国一体化算力网和“东数西算”工程的部署，我国各地智能计算中心加快布局，30余城市落地40余智算中心。

①目前，各地智算中心如雨后春笋快速建立，全国各省市均有所布局。根据国家信息中心与相关部门联合发布的《智能计算中心创新发展指南》，当前我国超过30个城市正在建设或提出建设智算中心，已建成的、在建中的和规划的中心数量已逾40家。一般智算中心的起步算力目标是100P，整体布局以东部地区为主，并逐渐向中西部地区拓展。



➔ 图15多地积极推动智算中心建设

资料来源：郑纬民院士学术报告《人工智能算力基础设施的设计、评测与优化》，网络

②根据ICPA 智算联盟统计，截至2023年3月，全国已投运的人工智能计算中心有近20个，在建设的人工智能计算中心超20个。2023至2024年，北京、哈尔滨、成都、广州、杭州、长春等多地持续发布智算中心相关招标公告。金额大多在亿元以上，招标单位具有国资背景。全国各地智算中心建设热度持续。各地方也结合本地产业特色，加快人工智能应用创新，聚合人工智能产业生态，例如武汉人工智能计算中心陆续孵化出紫东·太初、武汉·LuoJia等大模型，加速推动AI在多模态交互、遥感等领域的落地应用。

表2各地人工智能计算中心建设情况(不完全统计)

合作方	中心名称	地域	算力	建设情况
中科曙光	青岛“海之心”人工智能计算中心	山东省青岛市崂山区	混合精度	建设中
	长沙人工智能创新中心硅立方	湖南省长沙市	混合精度	已建成
	万达开先进计算中心	四川省达州市	混合精度	已建成
	宜昌先进计算中心一期	湖北省宜昌市	混合精度	已建成
	芜湖一体化智算中心	安徽省芜湖市	混合精度	建设中
	合肥先进计算中心“巢湖明月”	安徽省合肥市	混合精度	已建成
华为昇腾	北京昇腾人工智能计算中心	北京市门头沟区	一期100P FLOPS (短期500P, 远期1000P)	已建成
	中原人工智能计算中心	河南省郑州市	计划100P FLOPS	已建成
	未来人工智能计算中心	陕西省西安市雁塔区	一期规划300P FLOPS FP16	已建成
	成都人工智能计算中心	四川省成都市	300P FLOPS (最终1000P)	已建成
	武汉人工智能计算中心	湖北省武汉市东湖区	100P FLOPS	已建成
	重庆人工智能计算中心	重庆科学城	一期400P FLOPS	已建成
	长沙人工智能计算中心	湖南省长沙市高新区	200P FLOPS (25年达1000P)	已建成
	横琴人工智能超算中心	广东省珠海市横琴区	1.14Eops (完全建成4Eops)	已建成
	杭州人工智能计算中心	浙江省杭州市	40P FLOPS (后期100P)	已建成
	南京鲲鹏·昇腾人工智能计算中心	江苏省南京市	800P OpS	已建成
	济南人工智能计算中心	山东省济南市	/	已建成
	青岛人工智能计算中心	山东省青岛市	100P FLOPS	已建成
	天津人工智能计算中心	天津市河北区	300P FLOPS	已建成
	河北人工智能计算中心	河南省廊坊经济开发区	100P FLOPS	已建成
	大连人工智能计算中心	辽宁省大连市	100P FLOPS	建设中
	沈阳人工智能计算中心	辽宁省沈阳市	100P FLOPS (后期300P)	已建成
	中国-东盟人工智能计算中心	广西省南宁市	一期建设40P训练系统和1.4P推理系统	已建成
	深圳人工智能融合赋能中心	广东省深圳市龙岗区	接入20余万路视频资源	已建成

表2各地人工智能计算中心建设情况(不完全统计)

合作方	中心名称	地域	算力	建设情况
华为昇腾	广州人工智能公共算力中心	广东省广州市	99P FLOPS	已建成
百度	阳泉智算中心	陕西省阳泉市	计划100P FLOPS	已建成
	盐城智算中心	盐城市盐南高新区	200P	已建成
	百度沈阳元宇宙智算中心	沈阳市皇姑区	一期200P 总体500P	建设中
华海智汇 (华为控股子公司)	合肥人工智能计算中心	安徽合肥	100P FLOPS FP16	已建成
商汤科技	商汤人工智能计算中心	上海自贸区临港新片区	同时接入 850万路视频	已建成
腾讯	腾讯长三角人工智能超算中心	上海市松江区	1400P FLOPS	部分建成
	智慧产业长三角(合肥)智算中心	合肥高新区	/	已建成
上海交通大学等	太湖量子智算中心	江苏省无锡市	/	已建成
浪潮、寒武纪	南京智能计算中心	江苏省南京市	800P0pS	已建成
浪潮信息	淮海智算中心	安徽省宿州市	300P FLOPS	建设中
	青田元宇宙智算中心	浙江省青田县	每秒算力性能 超10亿亿次	建设中
	克拉玛依智算中心	新疆克拉玛依市	机柜数量超 1万个	已建成
寒武纪	昆山智算中心	江苏省昆山市	建成后峰值 智能算力不低于 500P0PS	建设中
阿里云	阿里云张北超级智算中心	河北省张家口市张北县	12EFLOPS	已建成
	阿里云华东智算中心	上海市金山区	在建	建设中
	阿里云乌兰察布超级智算中心	内蒙古乌兰察布市	3EFLOPS	部分建成
中科曙光、 华为	长沙5A级智算中心	湖南省长沙市	建成后算力规 模可达1024P	已建成
宁数科创	宁波人工智能超算中心	宁波市高新区	一期100P (FP16)+ 5P(FP64) 二期300P (FP16)+ 15P(FP64)	部分建成
福州市电子 信息集团	福州智能计算中心	福建省福州新区	一期105P, 总体400P	已建成
浪潮云、 中国电信 武汉分公司	武昌智算中心	武汉市武昌区	建成后100P	建设中

表2各地人工智能计算中心建设情况(不完全统计)

合作方	中心名称	地域	算力	建设情况
/	哈尔滨人工智能先进计算中心	哈尔滨	100P	已建成
/	北京数字经济算力中心	北京	建成后1000P以上	建设中
/	郑州人工智能计算中心	河南省郑州市	一期2000P 二期10000P, 建成后30000P	建设中
/	石景山智算中心	北京北重科技文化产业园	建成后610P	建设中
/	北京七星园数字经济产业智算中心	北京市丰台区	建成后2300P	建设中
/	华南数谷智算中心	韶关市武江区	一期16000P	建设中
/	博大数据深圳前海智算中心	深圳	一期40000P	已建成
/	中国移动智算中心(青岛)	青岛	建成后825P Flops, 参数网 络带宽达200GB	建设中

数据来源：至顶智库，通信世界全媒体，中国新闻网，澎湃新闻，政府网站

各地智算中心建设分两种模式，大多数采取“政府主导+企业承建”模式，政府结合业务需求与企业规模等因素进行综合考量、选取合适的承建单位。智算中心主要分为“企业自建模式”与“政府主导+企业承建”模式。智算中心具有高投入、对地方经济发展具有高影响等特点，因此30多座城市的智算中心建设项目大多数由政府主导，并紧密配合“东数西算”等建设指引的推进节奏，用于支持地方产业AI化、AI产业化以及智能化治理。中科曙光、华为、百度、商汤、腾讯、阿里、浪潮信息和寒武纪等众多厂商参与建设。

① 中科曙光率先提出5A级智算中心解决方案，拥有全面精度、算力融合、存算协同、领先液冷、生态兼容、建运一体等系列优势，其承建的智算中心在湖北、山东、江苏、安徽、浙江、广东和湖南等多地生根。

② 华为拥有雄厚财务基础、扎实的技术水平和全体系昇腾智算框架，成为众多智算中心的供应商。

③ 寒武纪不断深入参与智算中心建设，接连中标南京智能计算中心一期与二期项目，总预算金额超八亿。

④ 浪潮、天数智芯、燧源等企业也纷纷加入智算公共平台建设。

中科曙光致力打造新型智能算力基础设施，建立“5A级”核心优势。中科曙光作为新一代人工智能算力平台参与方，提出要打造具备“开放、融合、绿色、普惠、服务”能力的“5A级”智算基础设施，对智算中心的多方面性能提出了具体要求。中科曙光“5A级”智算中心为基础设施建设提供可参考的高质量范本，强力推动未来智算产业生态开放和协同发展，使智能算力基础设施迈入发展快车道。



图16中科曙光智算中心布局

资料来源：中科曙光

曙光形成“芯-端-云”全产业链布局，参与了多地智算中心的建设运营。其中长沙人工智能创新中心硅立方由长沙市政府与长沙高新区共同出资，长沙麓谷城市发展有限公司、中科曙光公司联合打造，按照中科曙光5A级智算中心标准进行建设，从底层架构开始，打通核心技术体系，既能形成多元融合算力供给，应对AI全场景需求，又能大幅提升兼容性，降低用户迁移和编译成本。硅立方整体采用了全球领先的浸没式相变液冷技术和冷板式冷却技术，可实现超低能耗、超高效率，打造绿色普惠算力，可为客户提供裸金属、虚拟机和容器等多种资源类型，满足各类算法、大模型等人工智能业务需求。其异构加速卡从部署到调通的周期普遍在5-7个研发人日，用户满意度较高。截止2024年4月份，长沙智算硅立方算力资源使用率达到60%，已与长沙千博信息技术有限公司、湖南迈曦软件有限责任公司、湖南科创信息技术股份有限公司等本地企业和高校建立了广泛合作关系，共同推进湖南国产AI算力生态的发展。



图17长沙人工智能创新中心硅立方

资料来源：中科曙光

北京昇腾人工智能计算中心成立，一期的算力规模实现100P,未来将实现1000P的算力规模，赋能金融、医疗、出行等多行业的大模型。2023年2月17日，北京昇腾人工智能计算中心正式点亮，首次采用市场化运作模式，打造产业发展新模式。该计算中心由北京市门头沟区政府联合中关村发展集团、华为公司携手打造，可为企业和科研单位等提供昇腾AI澎湃算力服务。该算力中心一期算力规模达到100P,未来，该智算中心将持续扩容：短期将实现500P的算力规模，长期将为企事业单位、科研院所提供1000P的普惠算力，助力人工智能产业蓬勃发展。



图18北京昇腾人工智能计算中心正式点亮及生态签约仪式

资料来源：华为计算

商汤承建临港AIDC智算中心，算力规模远超同业水平，可满足同时训练20个千亿参数量大模型。商汤AIDC计算中心占地超过13万平方米，一期机柜数量5000个，现已正式投入使用。中心通过部署2.7万块GPU，实现了超过4910 Petaflops的总算力供给，大幅超过业界绝大多数在200-300 Petaflops区间浮动的智算中心算力规模，通过完成行业十倍的规模提升跨越式提高了用户体验。在庞大规模的算力支撑下，商汤AIDC可以同时满足训练20个千亿参数量大模型的性能要求，算力可支持的单个最大模型的参数量超过万亿，能够充分满足各大厂商研发大模型的基础设施需求，支撑中国AI智算产业高速发展。商汤AIDC不仅规模庞大，同时达到了低碳节能的绿色环保发展要求。商汤通过采取各种能源优化措施、实施离心系统并部署工业冷却制冷剂，通过提高每摄氏度3%到5%的冷却效率，有效实现80%的降低幅度，年均PUE 优化至1.28，能耗平均水平比国内其他数据中心低约10%。AIDC是支撑商汤大模型的有力基础计算平台，同时也是向社会输送普惠算力的坚实底座。



图19上海临港AIDC人工智能计算中心

资料来源：商汤官网

二、智算产业蓬勃发展，产业生态逐步形成

2.1 智算中心建设迅速落地，多方主体积极参与

智算中心拥有大规模智能算力集群，具有高性能计算能力，能够处理海量数据，为AI计算提供了高效的并行计算能力，通常还集成了各种优化的AI算法、框架和开发工具，加速了模型开发和训练过程。随着生成式AI的发展，大模型时代百花争鸣，各地积极推动智能化转型升级。由于智算中心强大的计算能力、高效的能源利用、计算资源的灵活调度与分配、以及计算与数据的安全可靠，各地智算中心建设正迅速推进落地。

智算中心对于推动技术创新和区域产业升级具有深远影响。智算中心为各行各业提供了人工智能和智能化实施的算力基础设施，为气候模拟、药物研发、天文学研究、生物信息等领域的复杂科学和工程问题提供智算资源和技术支持，推动前沿科技发展。同时，除了能够为各行业提供定制化的AI服务外，智算中心的建立通常伴随着人工智能、大数据技术生态系统的形成，包括创新型企业、研究机构和教育机构，促进了技术交流和知识转移，成为区域技术生态系统建设的关键，吸引带动先进技术核心产业集聚发展，加速区域产业结构优化升级。

智算中心的建设和运营依托于多方参与，主要有企业厂商和政府等公共机构两大主体。建设方包括地方政府、高校和科研院所、云服务商、运营商、行业龙头企业等。建成后也需要专业的运营服务机构来进行日常的运营管理和维护，工作包括算力资源的分配、模型开发和应用技术的支持、服务质量的保障以及客户关系的维护等。

表3我国智能计算中心建设主体分类

建设主体 (主导建设的一方)		建设目标	代表案例
地方政府		提供普惠人工智能算力服务	长沙人工智能创新中心硅立方 万达开先进计算中心 合肥先进计算中心 天津市人工智能计算中心 北京人工智能公共算力平台
科研院所、高校		服务科研项目	复旦大学云上科研智算平台CFFF
企业	云服务商	提供数据存储、通用	阿里云乌兰察布智算中心
	运营商	AI计算与大模型应用等方面的商用服务，部分智算中心面向智能驾驶等特定行业领域提供AI算力服务	中国电信京津冀大数据智能算力中心
	行业龙头企业等		吉利星睿智算中心

资料来源：《2023人工智能2.0时代的公共智算服务发展指南》，网络

在智算中心的建设和发展中，政府相关机构扮演着至关重要的角色。政府不仅提供政策支持和资金资助，还负责制定行业标准和监管措施。在中国《新一代人工智能发展规划》明确提出，中国到2030年成为世界主要人工智能创新中心。该目标离不开智能算力基础设施的高质量建设与支持。《中国制造2025》战略中，重点提到了智能制造和数字化工厂升级，这也使得作为支撑智能制造的核心技术的智能算力受到广泛关注。当前我国地方政府在北京、深圳等多个城市积极推动人工智能产业园、智算中心的建设，形成了一批人工智能公共算力平台。同时通过提供税收优惠、资金补贴等政策支持，鼓励和吸引企业在智能算力领域进行投资和发展。同时，中国政府也联合高校、企业等制定人工智能和智能算力领域的国家标准和行业标准，以指导和规范行业发展。

参与智算中心的企业厂商主要是提供关键硬件和软件设备的制造商。在硬件领域，有专门生产GPU、ASIC、FPGA和NPU 高性能计算硬件的制造商，以及生产存储和网络硬件，提供固态硬盘和高性能服务器存储系统高速存储解决方案来支持大规模数据的快速传输和处理的企业。在软件领域，主要包括提供人工智能和大数据处理的软件工具和平台，如机器学习框架、数据处理分析、云计算和人工智能大模型等计算平台。这些厂商的产品和服务是智算中心运行的基础，他们通过提供先进的计算、存储和网络技术，支持智算中心处理复杂的AI 任务，满足不断增长的数据处理需求。

当前我国注重构建人工智能公共算力开放创新平台，为科技创新和产业智能化转型提供人工智能普惠算力。在智算平台建设中，沈阳人工智能计算中心、大连人工智能计算中心、北京昇腾人工智能计算中心、南京智能计算中心、武汉人工智能计算中心、成都智算中心、西安未来人工智能计算中心、广东智能科学与技术研究院、杭州之江实验室这9家平台获批建设国家新一代人工智能公共算力开放创新平台。这些公共算力平台链接了“中国算力网”，广泛联合国家级智算中心、超算中心及全国一体化算力网络枢纽节点，推动全国大型算力协同调度与高效计算。同时，国家也注重高性能计算中心与智算中心的异构融合发展，支持天津、长沙、成都、郑州等超算中心进一步筹建国家新一代人工智能公共算力开放创新平台，并部署建设国家超算互联网，构建具备互联网特征的超级融合算力基础设施，推动自主软硬件生态创新发展。这些举措对于培育国产化人工智能生态、促进创新孵化、算力普惠、产业聚合具有重要作用。

智算中心如火如荼的建设是时代经济发展的需要与技术进步的必然结果。对于企业而言，随着人工智能和大数据的快速发展，对高效计算资源的需求不断增长，智能算力市场存在巨大潜力和商机。各企业通过参与智能算力发展，能够推动自身的技术创新，帮助自身优化数据处理流程、提升产品和服务质量，从而提高整体运营效率。同时参与智能算力发展也为开发云计算服务、人工智能驱动解决方案等新业务提供技术支持，保持在竞争激烈的市场中的领先地位。对于政府而言，政府通过支持智能算力发展，提升国家在全球科技竞赛中的地位，加强技术创新能力。智能算力作为新兴产业的关键基础设施，能够带动高科技行业的发展，促进经济结构的优化和升级，在高技能领域够创造新的就业机会。对于公共事业，政府支持智能算力，以提高如医疗、交通、政务服务等领域公共服务的效率和质量。对于科研机构和教育机构，智能算力提供了强大的计算资源，支持科研机构进行复杂的科学研究和技术创新。同时，与智能算力相关的项目和研究为教育机构提供了实际的教学和培训平台，有助于面向未来的科技人才培养。

2.2 智算产业链正在形成

智能算力产业链是围绕人工智能技术的计算需求构建的一系列相关产业的集合。它包括为人工智能提供必要支持的各种硬件、软件、服务和应用，涵盖了从芯片制造到人工智能应用开发的全过程。当前，我国智算产业链正在形成，逐渐搭建起以提供算力硬件设备和基础设施为主的上游，以提供智能算力服务为主的中游，和聚焦智能算力应用的下游。日渐完备的智算产业链正以更高效的方式及集合支持着人工智能算法的研发，成为我国夺取未来经济发展新高地的重要支撑。

在产业链上游，主要涉及硬件设备和基础设施的生产和开发，包括GPU、ASIC等芯片制造、高性能服务器与存储设备制造、数据与计算中心建设等。这些硬件是智能算力产业链的物理基础，为AI应用提供了必要的计算和存储能力。在芯片领域上，华为公司开发的昇腾系列AI处理器，海光信息独立研发的深算系列DCU,均提供了面向行业大模型领域与其他应用场景的深度学习、模型优化、一站式开发等方面的解决方案。比如美国的英伟达开发的GPU 芯片，提供高性能、能效和高带宽连接，可用于各种配置，以满足不同的数据中心需求。在服务器和存储设备领域上，具有大量处理器和存储资源的高性能服务器等硬件设施构成与夯实了人工智能计算平台的物理基础。例如浪潮公司提供高性能、高扩展性的NF 服务器与AS存储系统的解决方案，能够支持企业级数据中心、云计算、大数据的需求。中科可控最新发布的高融合一体化算力平台—整机柜服务器 D9000 G5, 不仅将整机柜算力密度提升了4-8倍，同时集成了超大规模全闪存存储能力存储容量高达3.2P, IOPS 超过61万，有效解决了大模型训练中语料加载缓慢及大模型Checkpoint 管理的两大问题，实现了数据的快速读取与高效处理。在数据中心基础设施管理解决方案上，机房环境监控和冷却系统优化也是产业链上游的重要一环。作为智算的重要集成性支撑，数据与计算中心建设如火如荼，数据中心的建设规划了能源利用系统和冷却系统等，考虑了安全性和可扩展性，布局了机房、电力供应等物理基础设施，平衡了性能与能源消耗，将智算产业链上游的各个部分有机集合起来，并为中下游产业提供基础。在智算产业链上游与其他技术上游产业链互有交叉，比如与5G、大数据、云计算等，这也是当前ICT 行业发展中高度融合与集成的特点所在。随着智算技术本身与相关配套供应链的完善与成熟，智算产业上游日趋成熟与完备。

在产业链中游，主要云集多种形式的智算服务，包括云计算平台、边缘计算平台、人工智能服务平台、软件开发和集成服务。这些服务使得人工智能的开发和部署变得更加灵活和高效，为人工智能应用提供了必要的支持环境。在云计算平台领域，其扩展了智算提供了灵活性、可扩展性和高效性，支撑从数据存储到人工智能算法的运行等各个方面，可以降低企业在算力领域的成本。当前中国的云计算服务在世界领域具有极强的竞争力，如阿里云的云服务器ECS, 可以提供丰富应用场景下高效的计算资源，支持大规模人工智能和机器学习任务。在边缘计算平台和服务领域，在处理实时数据和降低云计算延迟方面，边缘计算减少了数据传输到云端的需要，从而显著降低了响应时间，适合应用于移动性强和地理分布广的场景，例如自动驾驶和物联网等领域的算力支持。百度BEC 平台对于自动驾驶过程中人工智能多模态分析与决策制定起到了关键作用，提高了自动驾驶的安全性与及时性。曙光网络公司在2023年发布了工业数智底座，旨在拉通工业系统与工业边缘计算，推动云边协同，加速矿山、电力、交通、钢铁等工业企业的数字化转型升级。

在人工智能服务平台，通过提供预建的模型、开发工具和APIs, 使客户轻松构建AI解决方案。除了阿里、百度、腾讯等巨头，中科曙光、神州数码等企业也提供多种灵活的人工智能服务，其他企业通过这些人工智能服务商，能更快速的享受智算行业带来的便利。在软件开发和集成服务领域，企业针对于智能算力应用的设计、实施和优化连接硬件能力和实际应用的关键环节，比如中科软、东软集团、用友网络等，提供包括信息集成系统、决策系统、风险管理系统等，通过融入智能算法，以提高数据处理和人工智能实地应用的效率。综上，智算产业链中游聚焦于智算服务，是连接上游硬件设备制造与具体应用场景的重要环节。

智能算力产业链的下游环节主要涉及智能算力的应用，这是将上游的硬件能力和中游的服务转化为实际价值的关键阶段。智能算力在下游的应用非常广泛，涵盖了工业制造业、医疗保健、金融服务、能源等多个行业和领域。在工业制造领域，通过基于图像处理技术和机器视觉进行产品质量检测，许多工厂采用自动化机器人在流水线进行快速决策和运动控制，比如比亚迪公司协助电池装配、焊接等工作等，富士康通过视觉系统进行产品质量检测和缺陷识别。在医疗领域，通过构建诊断模型深度学习X光、CT 扫描等影像资料，实现疾病诊断，也可以通过演算与分析加速新药研发和测试。腾讯的医学影像平台觅影提高医学影像的诊断准确率，辅助医生在肺癌、乳腺癌等早期诊断，复星医药也使用智能算法对大量的化合物和生物数据进行分析，以寻找新的药物候选物。在金融服务领域，可以提供交易决策辅助和风险分析管理。支付宝中就有识别异常交易和预防欺诈活动，提高风险管理的效率和准确性，国泰君安证券在其交易平台中应用智能算力分析市场数据来优化交易决策，提高交易效率和盈利能力。在能源领域，可以通过人工智能实现能源生产与分配的管理。国家电网公司利用大数据和AI技术进行电力需求预测、负荷管理、故障检测和电网运维优化，以提高电网的效率和可靠性。综上，智能算力产业链的下游覆盖各行各业，下游的应用是使智能算力真正赋能经济发展的最实际的部分。



图20智算产业链图谱

资料来源：上海人工智能研究院根据公开资料绘制

三、智算芯片是智算产业核心环节，也是全球科技巨头竞逐热点

3.1 智算芯片是中美科技博弈重要领域，国产智算芯片发展刻不容缓

3.1.1 智算芯片是智算产业的核心环节

智算芯片是智算产业的关键基础。智算产业服务于人工智能行业应用，而计算能力是人工智能算力的基础。智算芯片，特别是GPU、TPU 等专为AI计算设计的芯片，提供了强大的计算能力，使得大规模的神经网络训练和推理成为可能。智算芯片通过优化硬件设计和架构，可以大大提高AI算法的运行速度和处理能力，从而提高整个智算产业的效率和性能。智算芯片通常具有较高的能效比，这意味着它们可以在消耗较少能量的情况下完成大量的计算任务。这对于降低智算产业的能源消耗和运营成本非常重要。据统计，在目前市场主流的高算力AI服务器领域，GPU等核心芯片占服务器的成本占比高达80%以上。

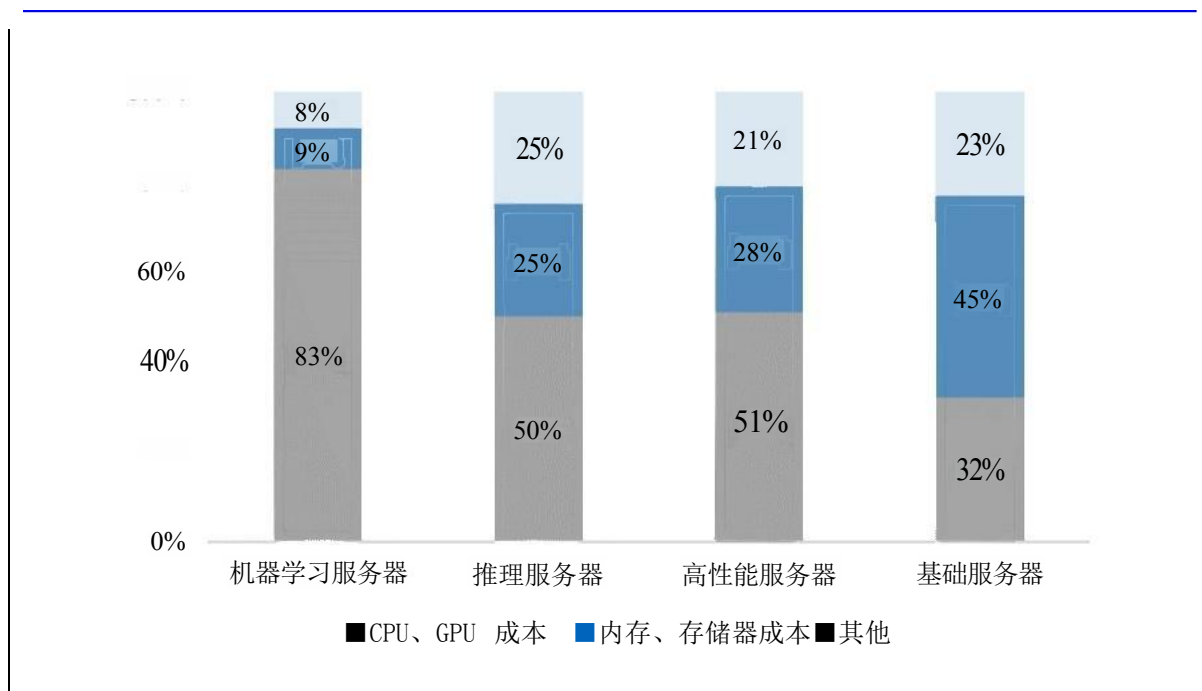


图21 服务器主要成本构成

资料来源：芯八哥

智算芯片的不断创新推动了整个智算产业的发展。新的芯片技术和架构不断出现，为AI算法和应用的创新提供了强大的驱动力。在不断挑战计算能力极限的同时，智算芯片也催生了新的应用场景和商业模式。其作为核心环节，通过不断创新，带动了智算产业的技术、产品和服务的创新，形成良性循环。

智算芯片成为产业融合的纽带。智算芯片融合了多种技术，包括微电子技术、计算机科学、人工智能等。这些技术的融合使得智算芯片能够在各种不同的应用场景中发挥重要作用，应用覆盖通信、交通、医疗、教育等各个行业，通过在各行各业应用中作为智能化转型的基础硬件设备核心，成为联接智能产业各个环节的纽带。

3.1.2 智算芯片成为中美科技博弈的焦点之一

智算芯片在全球范围内展现出巨大的市场和经济价值。随着人工智能、大数据等新兴技术的发展，由于智算芯片的核心技术地位，智算芯片的战略价值日益凸显，代表着高端技术和创新能力。全球每年的智算芯片市场规模达数百亿美元，凸显出巨大的市场和经济价值。根据Precedence Research统计，2022年全球AI芯片市场规模为168.6亿美元，预计到2032年将达到约2274.8亿美元，年复合增长率达到29.72%。



资料来源: Precedence Research

智算芯片成为大国科技博弈的焦点之一。智算芯片作为很多关键基础设施行业领域中的核心硬件组成部分，不仅关系到经济发展，也关系到国家安全。智算芯片的制造和应用形成了庞大的产业链，包括设计、制造、封测等环节，涉及到全球产业链的重要节点，智算芯片的关键技术和市场份额影响了全球科技格局。智算芯片的发展也关系到国际科技规则的制定和话语权的争夺。在这一领域的领先地位意味着能够主导相关标准和规范，影响全球科技治理。智算芯片领域的大国博弈不仅仅是技术的角逐，更关于未来科技规则的塑造和制定。

美国陆续推出限制性政策限制中国智算芯片的发展。智算芯片的经济和战略价值使其已经成为了中美两国竞争的重要领域。中美两国都在加大投入，力争在智算芯片领域取得领先地位。这种竞争不仅体现在技术层面，也体现在市场层面和政策层面。美国通过一系列的法案和政策，试图限制中国智算芯片的发展。美国政府通过实施严格的出口管制，限制向中国出口关键的智算芯片及相关技术，包括对先进制程设备、关键材料、设计软件等的出口限制，影响了中国芯片产业链的上游环节。美国商务部将中国的某些科技企业列入实体名单(Entity List)，禁止美国公司或使用美国技术的外国公司与这些企业进行交易，限制其获取先进的智算芯片技术和产品。国内AI芯片和半导体行业面临严峻的挑战，寒武纪、壁仞科技等公司就先后被美国商务部工业和安全局(BIS)以国家安全和外交利益为由列入了“实体清单”。通过一系列制裁措施，美方不断对中国智算芯片行业施加压力。

表4近年海外对华半导体管制措施(部分)

时间	管制主要内容
2018年4月	美国商务部发布公告，在未来7年内禁止中兴通讯向美国企业购买敏感产品。
2019-2020年	2019年5月，华为及68家附属关联公司被美国列入“实体名单”；2020年5月，BIS限制华为购买使用美国技术、软件设计制造的半导体；2020年8月，BIS在实体清单中新增38家华为附属公司，并修订外国制造直接产品规则，进一步限制华为使用基于美国软件/技术生产的半导体。
2020年12月	中芯国际被纳入实体名单，对用于 $\leq 10\text{nm}$ 技术节点的产品或技术，美国商务部采取“推定拒绝”的审批政策进行审核。
2022年7月	美国众议院通过《芯片与科学法案》，主要包括：(1)分5年提供527亿美元用于半导体制造激励计划、研发投资、税收抵免，其中美国芯片基金共500亿美元，390亿美元用于鼓励半导体制造企业，110亿美元补贴芯片研发；(2)法案授权在未来十年拨款2000亿美元增加关键领域科技研发的投资；(3)法案要求获得补贴的半导体企业未来10年内不得在中国大陆新建或扩建先进制程的半导体工厂。
2022年7月	美国半导体厂商收到美国商务部规定，要求不得向中国供应用于制造 $\leq 14\text{nm}$ 芯片的设备。
2022年8月	BIS公告美国准备对EDA等四项技术实行出口管制。
2022年8月	美国通知英伟达向中国和俄罗斯出口A100和H100芯片需新的许可证要求。
2022年10月	BIS对中国实施超级计算机芯片和包含此类芯片的计算机商品的禁令；对受到许可证要求限制的外国生产项目的范围扩大到实体名单上中国境内的28家现有实体；针对 $\leq 18\text{nm}$ 的DRAM、 ≥ 128 层的NAND存储芯片增加了新的许可证要求；限制美国人员在没有许可证的情况下支持中国某些半导体制造设施的研发和集成电路的制造；将包括长江存储、中国科学院大学等科研院校在内的31家实体列入未经核实名单(UVL)。
2023年3月	美国商务部将浪潮集团、龙芯中科等公司列入实体名单。
2023年10月	美国商务部将壁仞科技、摩尔线程等公司列入实体名单。

资料来源：美国商务部、公开资料

表5 2023年10月被新列入“实体清单”的13家中国企业

英文名称	中文名称
Beijing Biren Technology Development Co.,Ltd.	北京壁仞科技开发有限公司
Guangzhou Biren Integrated Circuit Co.,Ltd.	广州壁仞集成电路有限公司
Hangzhou Biren Technology Development Co.,Ltd.	杭州壁仞科技开发有限公司
Light Cloud(Hangzhou)Technology Co.,Ltd.	光线云(杭州)科技有限公司
Moore Thread Intelligent Technology (Beijing)Co.,Ltd.	摩尔线程智能科技(北京)有限责任公司
Moore Thread Intelligent Technology(Chengdu)Co.,Ltd.	摩尔线程智能科技(成都)有限责任公司
Moore Thread Intelligent Technology (Shanghai)Co.,Ltd	摩尔线程智能科技(上海)有限责任公司
Shanghai Biren Information Technology Co.,Ltd.	上海壁仞信息科技有限公司
Shanghai Biren Integrated CircuitCo.,Ltd.	上海壁仞集成电路有限公司
Shanghai Biren Intelligent Technology Co. ,	上海壁仞科技股份有限公司
Superburning Semiconductor (Nanjing)Co.,Ltd.	超燃半导体(南京)有限公司
Suzhou Xinyan Holdings Co.,Ltd.	苏州新岩控股有限公司 (现为上海新之砾企业发展有限公司)
Zhuhai Biren Integrated Circuit Co.,Ltd.	珠海壁仞集成电路有限公司

资料来源：美国商务部BIS

3.1.3 国产智算芯片发展刻不容缓

智算芯片是构筑算力的重要基石，对于推动人工智能、大数据、云计算等智能计算技术的发展具有重要意义。智算芯片高效的并行计算能力和优化的硬件架构使得复杂的人工智能算法得以迅速执行，推动了在图像识别、自然语言处理、语音识别等领域的前沿科技进展，也为科学家提供了强大的工具，在天文学、生物学、化学等领域推动了前沿科技的突破和创新。

中国智能算力未来发展空间巨大。根据IDC和浪潮信息发布的《2023-2024年中国人工智能计算能力发展评估报告》，预计到2027年通用算力规模将达到117.3 EFLOPS,智能算力规模达1117.4 EFLOPS; 2022-2027年期间，预计中国智能算力规模年复合增长率达33.9%, 同期通用算力规模年复合增长率为16.6%。



→图23中国智能算力规模及预测 (2020年-2027年)●

资料来源：《2023-2024年中国人工智能计算能力发展评估报告》

然而目前我国在智算芯片领域仍存在较大的差距。我国智算芯片领域大部分市场份额被国外厂商占据，特别是在高端芯片市场。根据IDC统计，2023年上半年，中国加速芯片的市场规模超过50万张，其中，GPU 卡占有90%的市场份额；从品牌角度看，中国本土AI芯片品牌出货超过5万张，**在整个市场中占比约10%**。高度依赖进口芯片可能使国家面临技术封锁的风险。发展国产智算芯片是实现技术自主可控、提升安全和核心技术竞争力的必由之路。

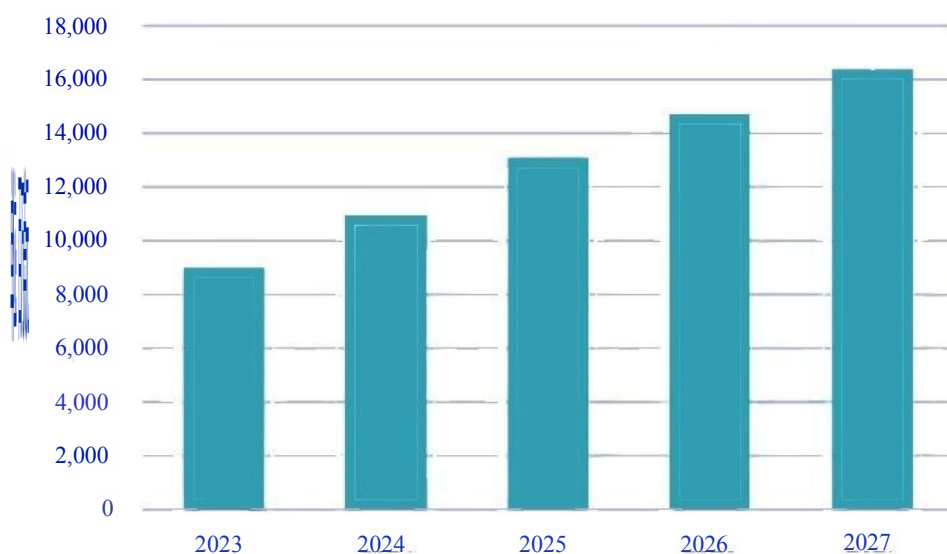


图24 中国加速计算服务器市场预测，2023-2027

资料来源: IDC中国

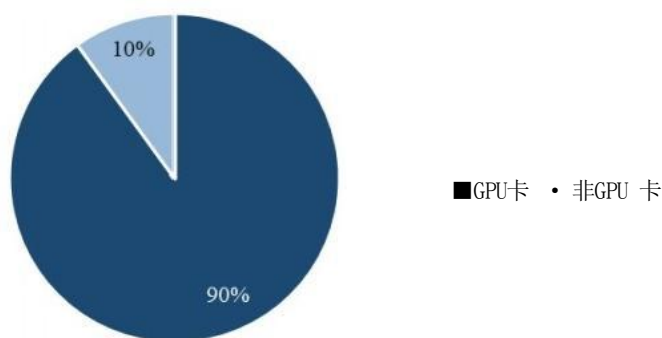


图25 中国人工智能芯片市场份额，2023H1

资料来源: IDC中国

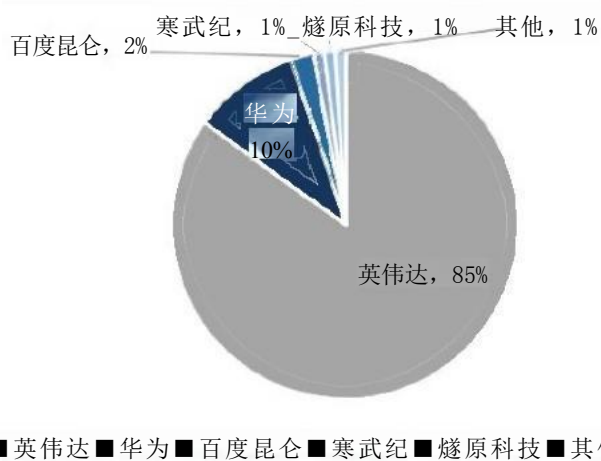


图26 2022年中国AI加速卡出货量占比

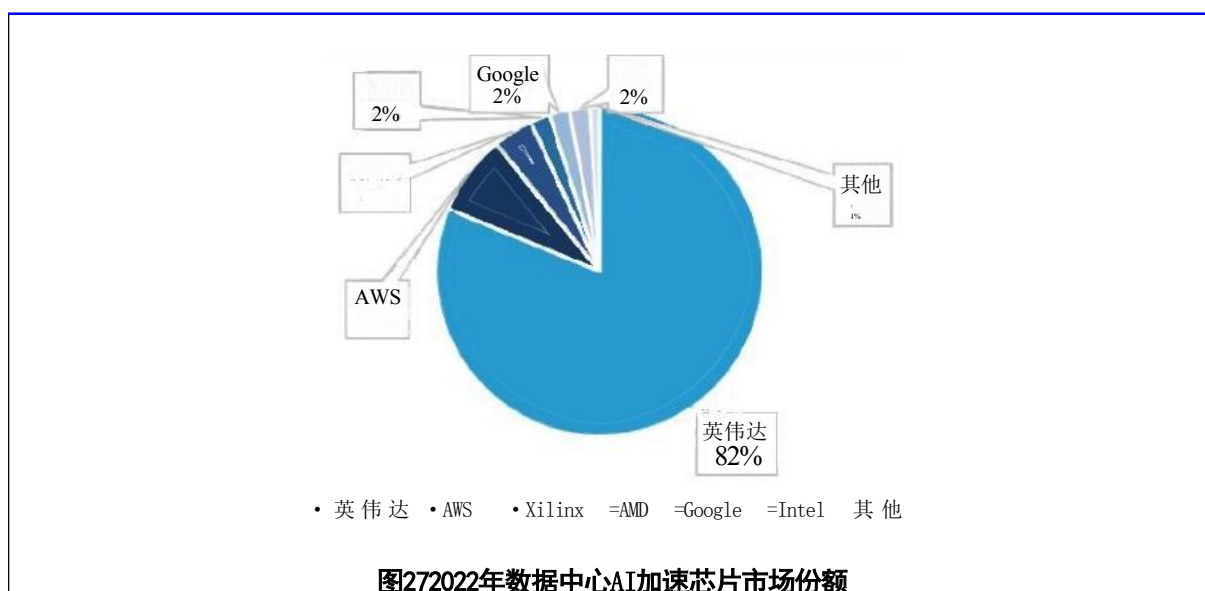
资料来源: 芯八哥

3.2海外：英伟达目前独大，AMD、微软、特斯拉、谷歌等开始发力

海外科技巨头英伟达、AMD、微软、特斯拉和谷歌等都在智算芯片领域进行了一系列布局，争夺市场份额和技术领导地位。他们通过不断推出新产品和技术，推动智算芯片的发展和普及，为人工智能和高性能计算等领域提供强大的支持。

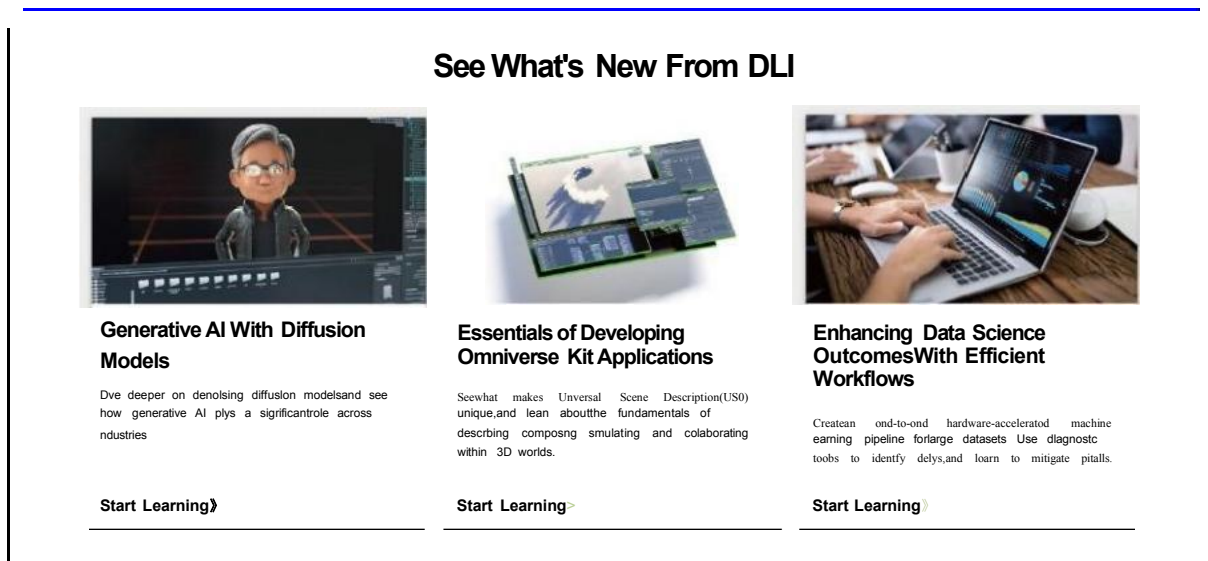
3.2.1英伟达目前在智算芯片领域独占鳌头

英伟达强大的GPU 产品线在深度学习和高性能计算等领域具有广泛的应用，被众多科研机构和企业采用，用于训练和部署人工智能模型，在AI和数据中心市场中占据主导地位。英伟达在发展其智算芯片生态方面采取了多种策略，成功地构建了一个强大且持续发展的智算芯片生态系统，促进了AI技术的广泛应用和创新。



资料来源：LIFTR INSIGHTS, 公开资料

在硬件创新方面，英伟达不断推出先进的GPU 和AI专用处理器，如Ampere 和Hopper 架构的GPU，以及Grace CPU, 针对AI和高性能计算任务进行了优化；在平台和软件生态方面，英伟达通过其CUDA 平台为开发者提供了一个强大的编程模型和工具集。CUDA 使得开发者能够有效地利用GPU 进行并行计算。英伟达还提供了一系列的软件库和框架，如cuDNN、TensorRT等，来支持深度学习和机器学习的开发和部署；在合作伙伴方面，英伟达与多个行业领导者建立合作伙伴关系，包括云服务提供商、电脑制造商和软件开发商。通过这些合作，英伟达能够将其产品和技术集成到广泛的行业解决方案中；在开源社区建设方面，英伟达积极参与开源项目，并为AI和GPU 计算社区提供支持。通过开源贡献，英伟达促进了技术共享和创新；在教育和研究支持方面，英伟达通过提供教育资源、培训课程和研究资助来支持学术界和研究社区。例如，英伟达的Deep Learning Institute(DLI)提供各种在线课程和研讨会。英伟达针对不同行业(如汽车、医疗、金融)提供定制化的AI解决方案，从而推动这些领域的技术进步。通过这些多元化的策略，英伟达成功地构建了一个强大且持续发展的智算芯片生态系统，促进了AI技术的广泛应用和创新。



→ 图 2.8 英伟达Deep Learning Institute

资料来源：英伟达官网

3.2.2 AMD在智算芯片领域不断发力

AMD 作为英伟达的主要竞争对手，也在智算芯片领域不断发力。AMD 的GPU 产品在性能和能效方面与英伟达展开激烈竞争，并且也在深度学习和高性能计算等领域取得了一定的市场份额。

AMD具备创新和多样化的产品线，推出了Ryzen和EPYC系列处理器，在多核心性能方面表现突出，适用于AI和机器学习任务。通过收购赛灵思(Xilinx), AMD 加强了在FPGA 和自适应计算领域的竞争力。在开发者支持方面，AMD 通过提供优化的软件工具包和开发者资源，如ROCm(Radeon Open Compute Platform)，支持开发者更好地利用其硬件进行AI开发，并与开源社区合作，提供对主流机器学习框架的支持，以确保其处理器和GPU能够与这些框架无缝集成。在合作伙伴关系方面，AMD通过与主要的OEM合作伙伴(如戴尔、惠普、联想)合作，将其处理器集成到企业级服务器和工作站中。与云服务提供商的合作，如亚马逊AWS、微软Azure等，使得AMD 的处理器成为云计算环境中的重要选择。在市场布局方面，AMD 积极推广其EPYC 服务器处理器系列，特别是在数据中心和云计算市场，以提供高性能的计算能力，特别适用于AI和高性能计算任务。



图2.9 AMD ROCm TM6开放式软件堆栈

资料来源：AMD官网

3.2.3 微软自研芯片垂直整合Azure 软硬件

微软在智算芯片领域的布局主要体现在其云计算服务Azure上。微软发布了首款自家研发的人工智能芯片Microsoft Azure Maia 100加速器，以及应用于云端软件服务的芯片Microsoft Azure Cobalt CPU。这两款芯片将由台积电代工，采用5nm 制程技术，预计将于明年导入微软Azure 数据中心，支持OpenAI、Copilot 等服务。

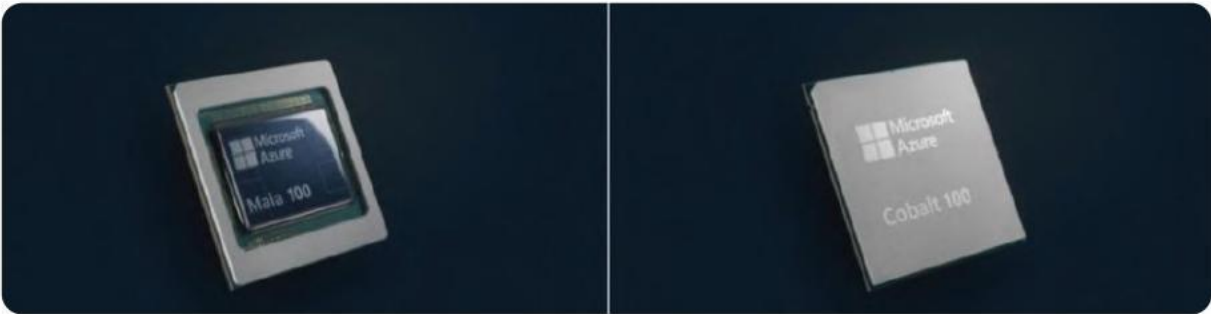


图30 MicrosoftAzure Maia 100和Cobalt 100

资料来源：公众号半导体行业观察

Maia 100 Specifications							
MSFP13 Matrix (OCP MXFP4)	MSFP17 Matrix (OCP MXint8)	BF16 Matrix	BF16 SIMD	L1 Cache	L2 Cache	HBM 3 Capacity	Memory Bandwidth (GB/s)
3200 TFLOPS	1600 TFLOPS	800 TFLOPS	49 TFLOPS	140MB	448 MB	64GB	1600
Maia 100 Specifications							
Backend Network Unidirectional	Gbps	SerDes Speed	PCIe	TDP	Node	Transistor Count	Die Size
600 GB/S	4800	112G	Gen 5x8	860W	TSMC N5	105 Billion	820mm2

图31 MicrosoftAzure Maia 100和Cobalt 100参数

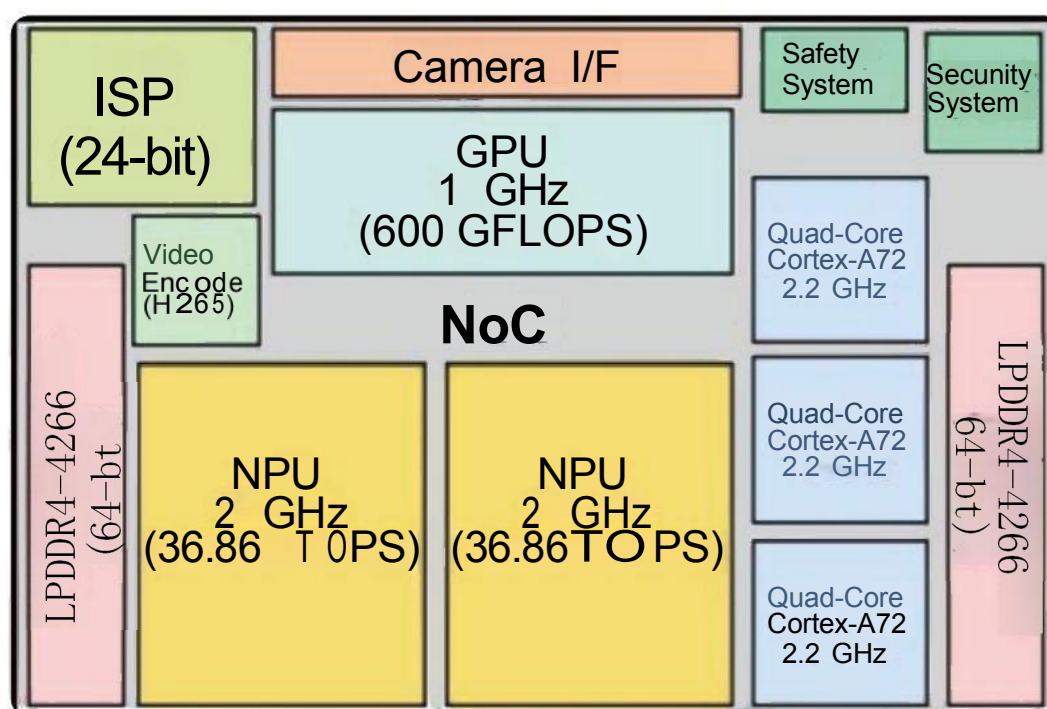
资料来源：36Kr

微软在智算芯片的生态构建上，更侧重于整合AI技术到其广泛的服务和平台中，而不仅仅是硬件产品的开发，从而使微软能够在云计算和AI服务市场中保持竞争力。微软通过提供一系列工具和平台，如Azure Machine Learning、ONNX(开放神经网络交换格式)等，支持AI的开发和部署。Windows 和Office等核心产品也整合了AI功能，如智能助手Cortana和自然语言处理技术。

3.2.4 特斯拉推出FSD 芯片以最大化其自动驾驶技术潜力

特斯拉作为全球领先的电动汽车制造商，也在智算芯片领域进行布局，主要专注于自动驾驶技术的需求。特斯拉推出了名为“FSD”的自动驾驶芯片，该芯片基于特斯拉自己的神经网络架构，用于实现高级别的自动驾驶功能。

特斯拉自2019年4月发布的Hardware 3.0开始采用其自研芯片FSD, 该芯片采用三星的14纳米工艺，2个神经网络处理器运行在2.2GHz 频率下芯片算力达到72TOPS。



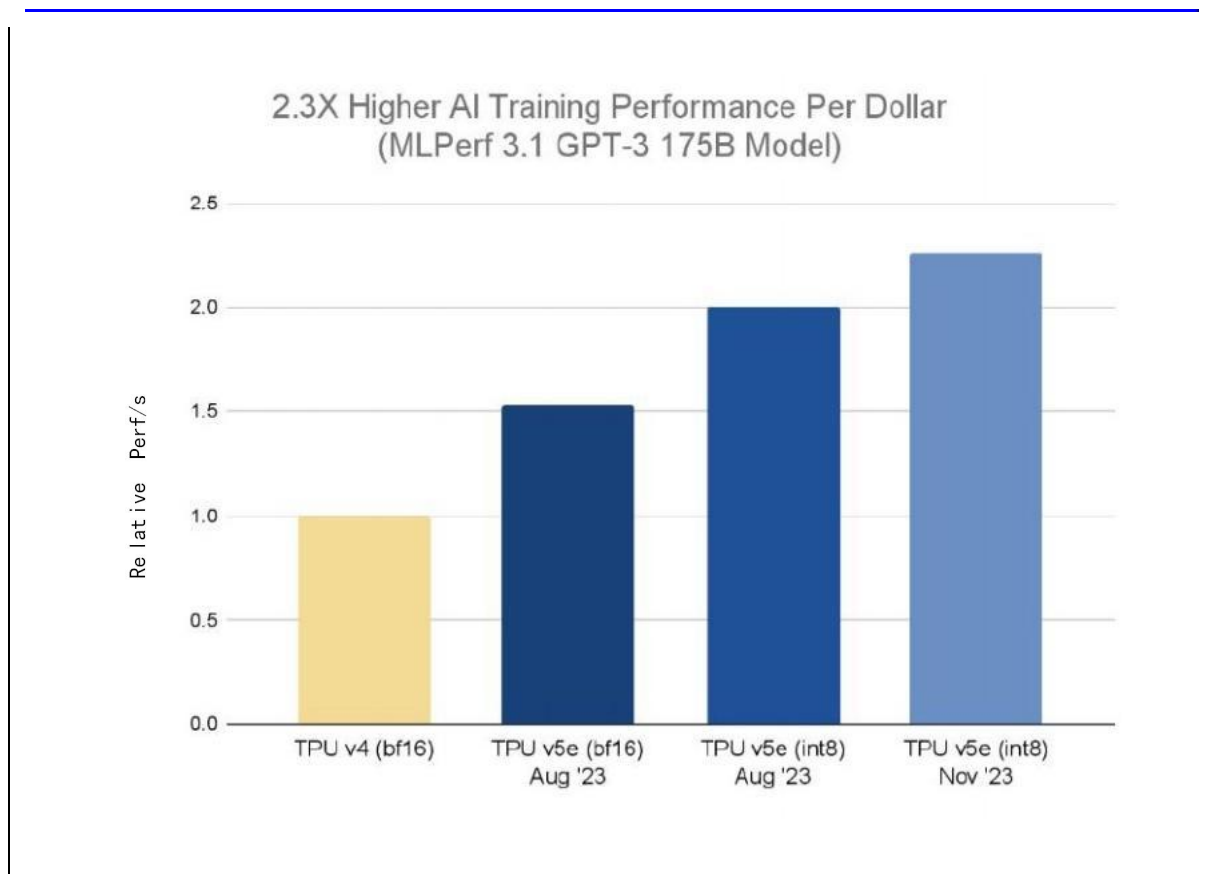
→ 图32 特斯拉FSD芯片架构简图 ●

资料来源：36Kr

特斯拉的智算芯片与其自动驾驶软件紧密集成，确保了软硬件之间的高效协同工作。特斯拉利用其车队收集大量数据，用于训练和改进其自动驾驶算法。通过实时数据的持续分析和学习，特斯拉不断优化其智算芯片的性能和效率，不仅在其新车型中标配了这些智算芯片，还提供旧车型的升级服务，以安装最新的自动驾驶硬件和软件。同时，特斯拉还积极推进Dojo超算集群的研发及建设进程，旨在更好地推动FSD算法升级迭代。

3.2.5 谷歌TPU 持续迭代性能升级

谷歌在AI领域的布局主要体现在其TensorFlow 框架和TPU（张量处理单元）上。TPU 是谷歌定制开发的应用专用集成电路(ASIC),用于加速机器学习工作负载,已经在多个谷歌的产品和服务中得到应用,如搜索引擎、语音识别和图像识别等。谷歌TPU 已经过数次迭代,2023年8月发布了第五代TPU v5e,专为提升大中型模型的训练、推理性能以及成本效益所设计,与TPU v4相比大型语言模型提供的训练性能提高了2倍、推理性能提高了2.5倍。



→ 图33 TPU v5e相对TPU v4的训练性能提升

资料来源: Google Cloud官网

谷歌通过其云平台Google Cloud提供TPU 加速的机器学习服务。客户可以利用TPUs 的强大计算能力,将其作为Google Cloud上的可伸缩计算资源使用,而无需直接投资于硬件。Cloud TPU能够针对各种AI工作负载(涵盖训练、微调和推理)进行经济高效的扩缩,提供了多种功能,可加速AI框架(包括PyTorch、JAX和 TensorFlow) 上的工作负载。客户还可以在全托管式AI 平台Vertex AI 中利用Cloud TPU。谷歌不仅在硬件上实现了创新,还构建了一个围绕TPUs 的综合性AI生态系统,包括云服务、软件工具和广泛的应用领域。

3.3国内：华为、海光信息等科技大厂积极布局，寒武纪、天数智芯等创业型企业纷纷兴起

近年来，我国的一些企业和研究机构开始重视智算芯片的研发，并取得了一定的进展。海光信息等IC设计企业通过技术创新，已独立研发出多款性能优异的高端通用智算芯片，并在金融、运营商、能源等关键行业规模化应用。华为、百度、阿里巴巴等互联网巨头公司已加快AI芯片的自研步伐，并在各自的云平台上率先试用。一些初创公司则从芯片的底层架构设计入手，试图通过“弯道超车”抓住机遇。

3.3.1华为打造全栈自主AI基础软硬件

华为推出了基于昇腾系列处理器构建的全栈AI计算基础设施及应用，基于昇腾系列AI处理器，通过模块、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，涵盖数据中心解决方案、智能边缘解决方案，覆盖深度学习领域推理和训练全流程。

在基础硬件方面，华为基于昇腾Ascend 系列芯片推出Atlas系列产品，提供AI训练、推理卡及训练服务器。在软件方面，华为构建了全栈全生命周期的模型开发工具链，通过全面的AI工具和服务，为业务智能快速创新赋能。在芯片使能驱动层推出异构计算架构CANN， 并构建了自主开源全场景AI框架MindSpore 以及一系列应用使能工具，包括行业SDK 和应用解决方案MindX SDK、AI全流程开发工具链MindStudio、AI开发平台ModelArts、优选模型库ModelZoo、深度学习使能MindX DL、智能边缘使能MindXEdge等，赋能能源、金融、公共、交通、电信、制造、教育等行业应用。



图34昇腾全栈AI软硬件平台
资料来源：昇腾官网

华为推出的异构计算架构CANN(Compute Architecture for Neural Networks)通过提供多层次的编程接口，以全场景、低门槛、高性能的优势，支持用户快速构建基于Ascend 平台的AI应用和业务。CANN 提供了一套标准的AscendCL 编程接口，能够对应用程序开发者屏蔽底层多种芯片差异，并且支持多框架，支持用户在昇腾芯片上快速部署神经网络业务。CANN 通过高性能编译引擎、执行引擎、调优引擎和预置高性能算子库，支撑客户快速部署神经网络业务、降低部署成本并最大程度发挥昇腾计算能力。在基础服务方面，CANN 还包含虚拟化、媒体处理、集合通信服务等基础服务，方便用户扩展及集成。

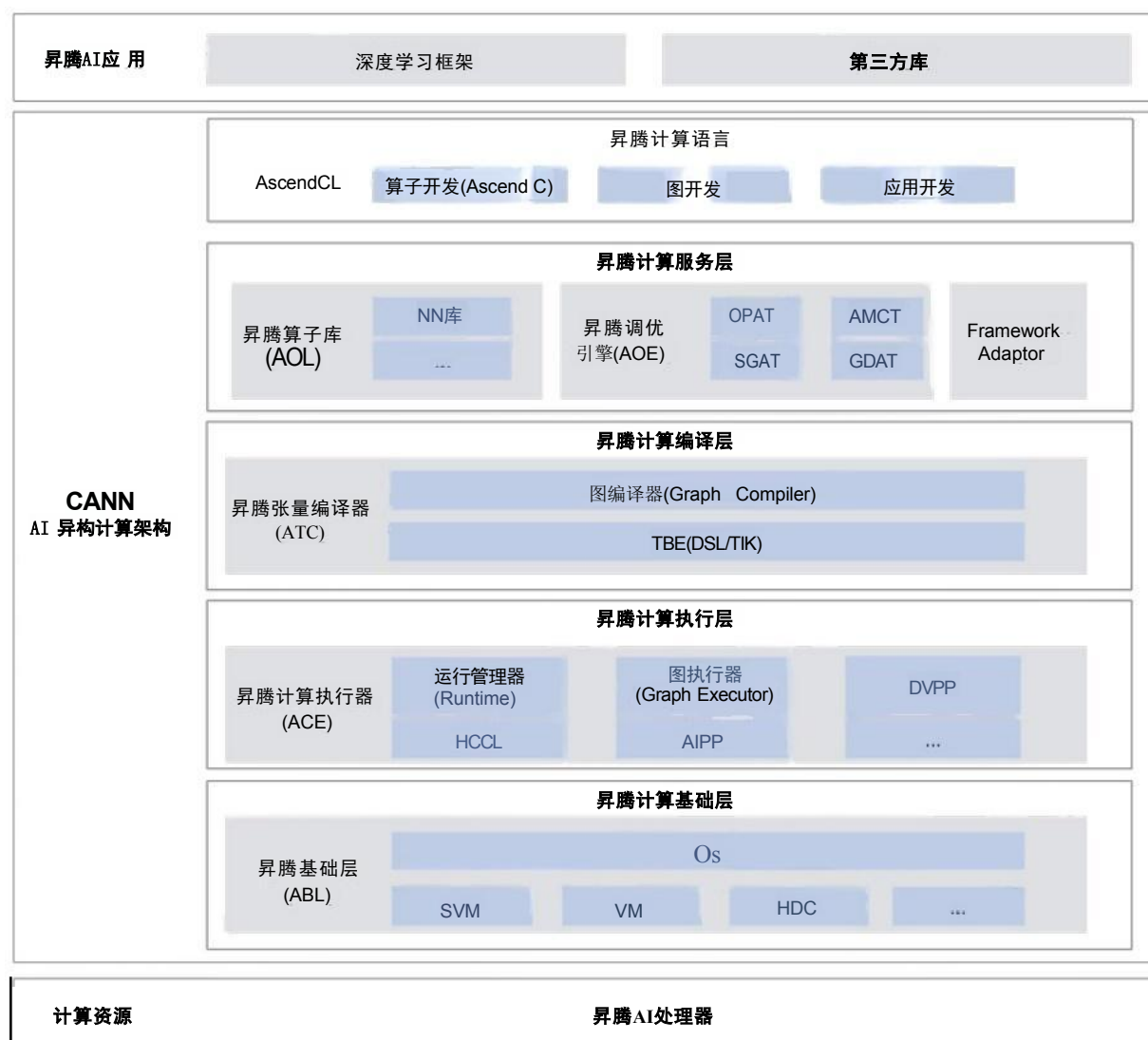


图35 CANN逻辑架构

资料来源：昇腾官网

华为自主构建的全场景深度学习框架Mindspore，旨在实现易开发、高效执行、全场景覆盖三大目标，同时支持端（手机与IoT设备）、边（基站与路由设备）、云（服务器）场景的不同系列硬件，包括昇腾系列产品、英伟达NVIDIA系列产品、Arm系列的高通骁龙、华为麒麟的芯片等系列产品。

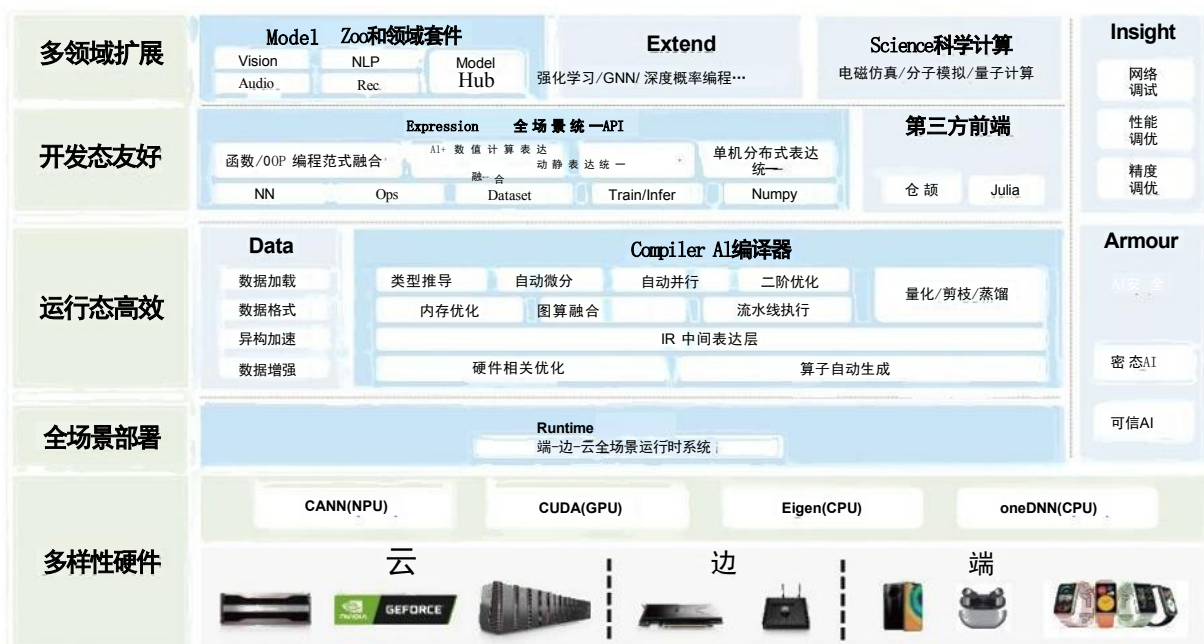


图36昇思MindSpore整体架构

资料来源：昇思社区

AI开发平台ModelArts 通过端到端模型生产线，为开发者提供高效开发、调试和调优大模型应用和场景化应用。MLOps 高效迭代AI模型，持续提升精度。平台提供AI加速套件，支持数据加速、训练加速和推理加速，支持分布式高效训练和推理，支持大规模异构集群及调度管理，并提供全流程云化昇腾迁移工具链，支持用户A 业务全栈国产化，帮助客户实现AI高效迁移。



图37 ModelArts AI开发平台

资料来源：华为云

在产业生态方面，昇腾计算围绕昇腾计算机软硬件体系，汇聚了包括原始设备制造商OEM、原始设计制造商ODM、独立硬件开发商IHV、咨询与解决方案集成商C&SI、独立软件开发商ISV、云服务提供商XaaS在内的产业合作伙伴，以及围绕昇腾相关产品对外提供服务交付的服务类伙伴、提供培训服务的人才联盟伙伴、提供投融资和运营服务的投融资运营伙伴，以及立足人才根本的高校和开发者，共同推动AI在各行各业的发展和繁荣。

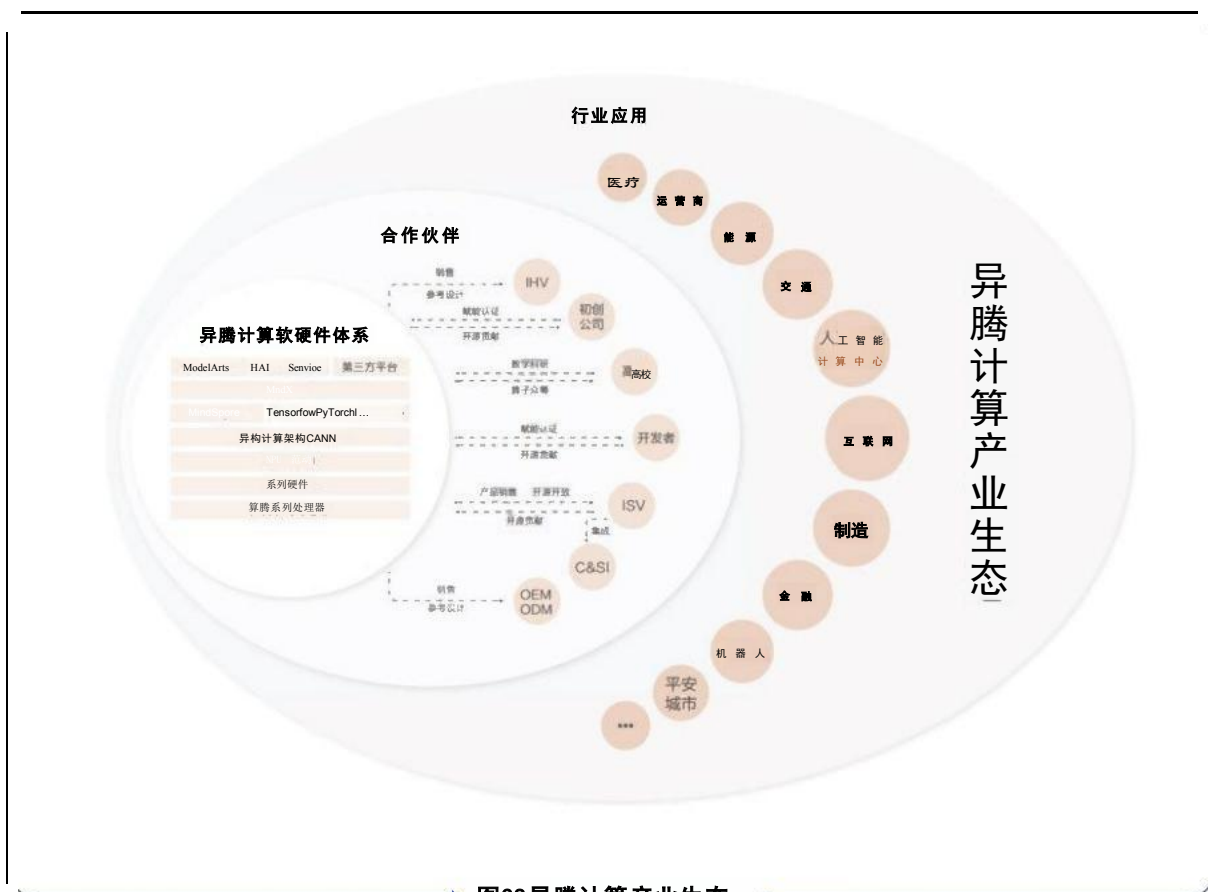


图38昇腾计算产业生态

来源：《昇腾计算产业发展白皮书》

3.3.2 海光DCU 依托开放生态加速推广

海光信息成立于2014年，主要从事高端处理器、加速器等计算芯片产品和系统的研发，是我国先进微处理器龙头。公司持续进行技术创新和演进，建立了完善的高端处理器的研发环境和流程，并不断提高产品性能、丰富产品功能，成功完成了多代产品的独立研发和商业化落地。

海光信息推出的DCU 产品，采用GPGPU 路线，具备全精度算力，性能强悍、生态完善。深算一号、深算二号两代产品已实现了在人工智能、大数据处理、商业计算等领域的规模化应用，可以用于大模型的训练和推理，客户覆盖了包括智算中心“新基建”、互联网、金融、运营商等行业。进一步地，海光与百度、阿里等头部互联网厂商推出了联合方案，打造了全国产软硬件一体全栈AI基础设施，形成了多个标杆案例。

海光信息DCU提供自主开放的完整软件栈，包括“DTK(DCU Toolkit)”、开发工具链、模型仓库等，其完全兼容“CUDA”、“ROCm”生态，支持TensorFlow、Pytorch 和PaddlePaddle 等主流深度学习框架与主流应用软件。依托开放式生态，海光信息构建了拥有完善的层次化软件栈的统一底层硬件驱动平台，其能够适配不同API接口和编译器，并支持常见的函数库、AI算法与框架等。实现了LLaMa、GPT、Bloom、ChatGLM、悟道、紫东太初等大模型的全面应用，DCU 也已与国内包括文心一言、通义千问等大模型全面适配，性能可以达到国内领先水平。



图39海光完善软件栈支持

资料来源：海光官网

3.3.3 寒武纪云边端协同发展

寒武纪公司自成立以来一直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，让机器更好地理解和服务人类。公司的主营业务是应用于各类云服务器、边缘计算设备、终端设备中人工智能核心芯片的研发、设计和销售，以及为客户提供丰富的芯片产品。

2022年，公司基于思元370云端智能芯片，推出了新款智能加速卡MLU370-X8/M8；推出了训练整机玄思1001智能加速器 (MLU-X1001)。与市场主流同尺寸GPU 相比，思元370系列加速卡在实测性能和能效方面表现出一定优势，尤其在视觉、语音、自然语言模型训练等场景的能效表现较为出色，不同尺寸的加速卡，充分满足不同场景下的客户需求。思元370实测算力性能能与主流GPU比肩。且思元370芯片支持LPDDR5 内存，高带宽且低功耗，可在板卡有限的功耗范围内给AI 芯片分配更多的能源，输出更高的算力。

在软件层面，寒武纪推出Cambricon Neuware基础系统软件，为云边端全系列智能芯片与处理器产品提供统一的软件开发工具链，打破了不同场景之间的软件开发壁垒，兼具高性能、灵活性和可扩展性的优势，可让同一人工智能应用程序便捷高效地运行在公司云边端系列化芯片与处理器产品之上，支持程序员实现跨云边端硬件平台的人工智能应用开发，以“一处开发、处处运行”的模式大幅提升跨平台开发效率和部署速度。Cambricon Neuware支持各主流人工智能编程框架，包括TensorFlow、PyTorch、Caffe、MXNet等，支持Paddle等国产人工智能编程框架，并同时支持训练和推理平台。开发者可直接基于主流编程框架的API为寒武纪的云端、边缘端、终端各款智能芯片和处理器产品方便地编写应用。

表6寒武纪主要产品

产品线	产品类型	寒武纪主要产品	推出时间
云端产品线	云端智能芯片及加速卡	思元100 (MLU100) 芯片及云端智能加速卡	2018年
		思元270 (MLU270) 芯片及云端智能加速卡	2019年
		思元290 (MLU290) 芯片及云端智能加速卡	2020年
		思元370 (MLU370) 芯片及云端智能加速卡	2021年、2022年
	训练整机	玄思1000智能加速器	2020年
		玄思1001智能加速器	2022年
边缘产品线	边缘智能芯片及加速卡	思元220 (MLU220) 芯片及边缘智能加速卡	2019年
IP授权及软件	终端智能处理器IP	寒武纪1A处理器	2016年
		寒武纪1H处理器	2017年
		寒武纪1M处理器	2018年
	基础系统软件平台	寒武纪基础软件开发平台 (适用于公司所有芯片与处理器产品)	持续研发和升级，以适配新的芯片

资料来源：寒武纪公司年报

3.3.4 天数智芯立足高性能通用GPU 加速构建AI自主生态

天数智芯是中国第一家通用GPU 高端芯片及超级算力系统提供商，致力于开发自主可控、国际领先的通用、标准、高性能的通用GPU 芯片，构建起全栈式软硬件体系，打造高端算力解决方案。天数智芯于2021年3月发布首款通用GPU 产品——天垓100，实现了国内通用GPU 从0到1的突破，2022年12月发布第二款通用GPU——推理产品智铠100，2023年公司推出了第三款产品“天垓150”。天垓系列产品采用完全自主研发的高性能 GPGPU IP核心、7纳米先进制程制造和 CoWoS 2.5D先进封装工艺，兼容CUDA生态，并兼容市场主流生态，可在云端训练、高性能计算和推理等应用领域对主流英伟达产品做到无缝切换。公司自研系统架构、指令集、核心算子、软件栈等，确保核心知识产权，保障安全可靠、独立演进。

天数智芯产品已经全面支持大模型预训练、微调及推理应用，协助国内首批客户完成了大模型在垂直领域的开发及部署。基于天垓100的自主算力集群已率先完成国内700亿参数大模型预训练，收敛精度达到国际主流产品水平且稳定性更高，成为率先支持大模型完整训练的国产通用GPU 产品，已搭建200P 自主算力集群，正在开展北京智源研究院700亿参数大模型预训练。此外，天垓100支持各垂直行业大模型训练微调，可支持全量参数微调、Lora快速微调、提示词微调等所有训练微调方法。智铠100已支持千亿级参数大模型推理，单机语言对话性能可达400token/s。

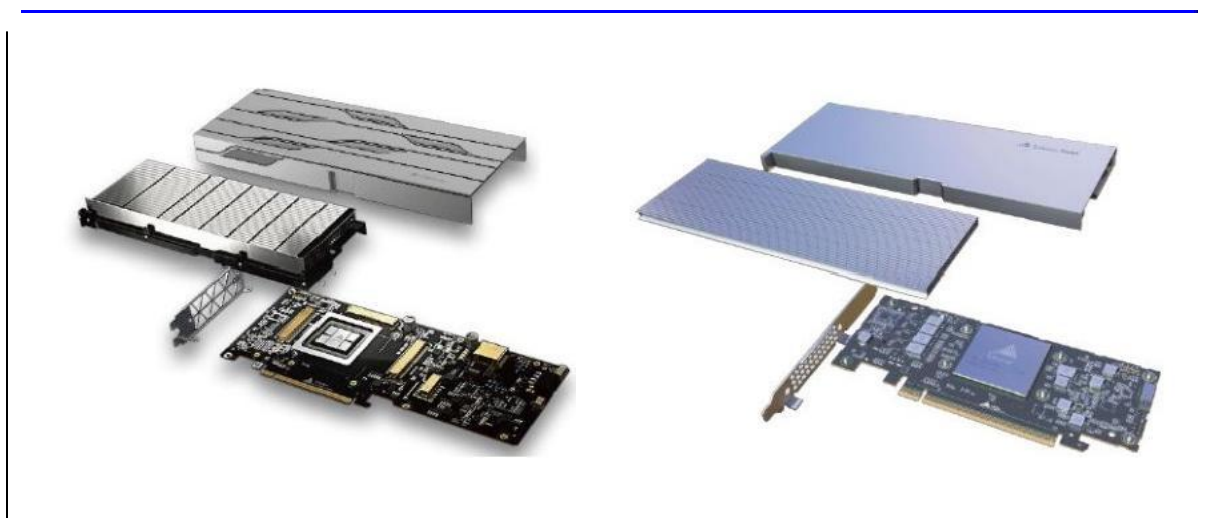


图40天数智芯GPU 产品(左：天垓100;右：智铠100)

资料来源：天数智芯

天数智芯通用GPU 产品支持国内外主流AI生态和各种深度学习框架，已支撑百余个客户在人工智能领域进行超过四百个不同种类模型训练，支持传统数学运算、加解密及数字信号处理等领域，获得了客户高度认可，在智慧医疗、运营商、大专院校、智慧教育等20余个行业实现落地应用，形成“应用覆盖广，开发易迁移，性能可预期”口碑，有效解决客户安全、替代难题，实现对新算法、新模型、新应用的快速支持。

3.3.5 瀚博半导体芯片产品聚焦AI加速和视频处理

瀚博半导体自主开发出核心IP和两代GPU芯片，进而衍生出AI、渲染、视频三大产品线，推出了载天系列AI加速卡、载天系列视频加速卡和南禹系列数据中心显卡。2022年，瀚博量产了基于第一代GPU芯片的智算中心通用智能AI加速卡，2023年4月，公司又量产了基于第二代全功能GPU芯片的渲染加速卡产品系列，在数字孪生、工业软件、云手机、云电脑等云渲染应用场景中展示了卓越的性能。

为提升产品的易用性，通过构建瀚博VUCA软件架构以及与生态和系统的兼容适配，不断优化用户体验，降低开发者使用的技术门槛。瀚博面向开发者提供一套完整的SDK(Vaststream)接口，并提供详尽的文档和示例代码，以帮助开发者更快速地上手和深入了解芯片的功能以满足开发者对芯片功能的灵活应用需求。同时为开发者与企业客户提供一整套方便易用的软件工具，用于实现云原生、集群监控、模型编译等能力。

瀚博统一计算架构VUCA(Vastai Unifined Computing Architecture),基于自研高性能SV100和SG100系列GPU芯片，搭载流媒体、AI推理和图形处理三大加速引擎，提供Vaststream统一API接口及极简业务应用开发套件，支持丰富的第三方标准库，并且配套开发、调测、监控工具链，帮助客户降低软件迁移成本和管理各类业务应用。

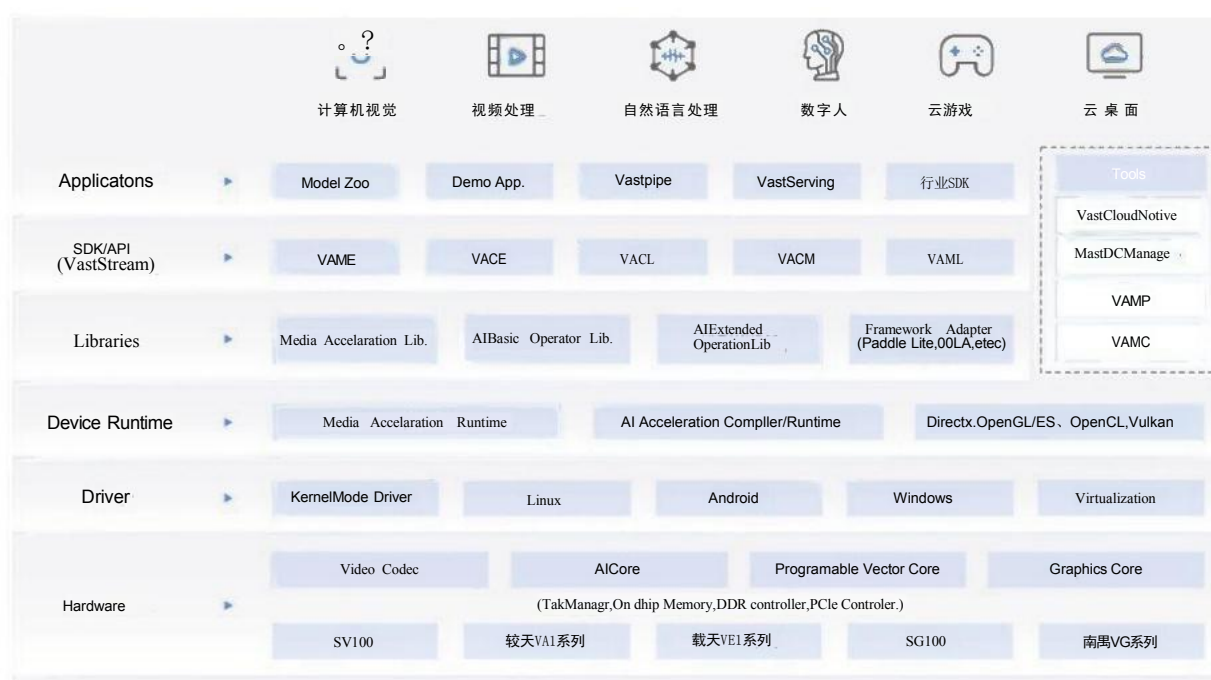


图41 VUCA 软件架构

资料来源：瀚博半导体

瀚博VastStream软件栈基于瀚博加速卡硬件，配套驱动，提供SDK封装，支持主流深度学习及视频开源框架，实现包括智能视频分析、云转码在内的用户应用生态。

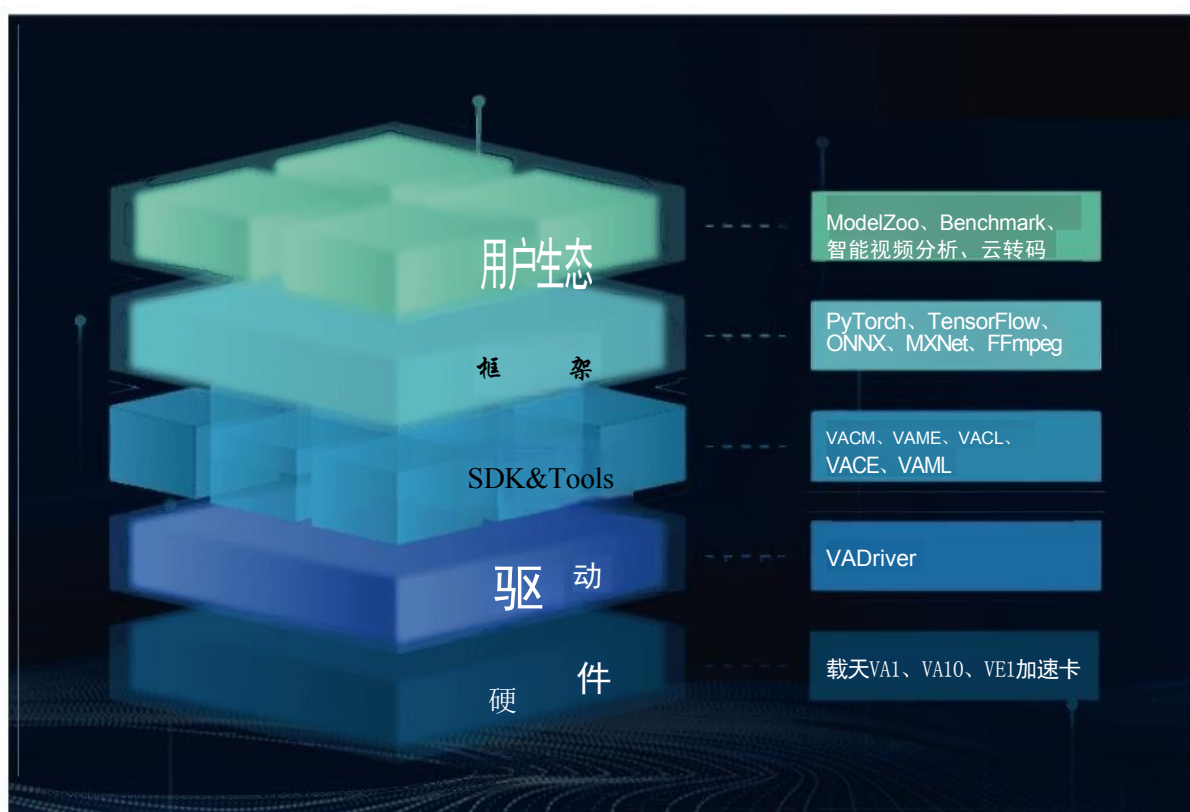


图42 VastStream软件栈

资料来源：瀚博半导体

为验证瀚博SV100芯片超强的AI算力与瀚博VUCA 软件的易用性，瀚博联合生态合作伙伴，结合需求和产品特点推出了一系列端到端场景应用解决方案，涵盖了多个领域，为各行各业提供了全面的技术支持。方案涵盖互联网、视频直播、车路协同、视频会议、工业质检、安防等多样化场景需求。针对互联网直播行业，基于瀚博半导体VA系列产品，为互联网直播平台提供了高效的视频转码和AI人脸增强解决方案，助力实现更流畅、高质量的直播体验；针对视频会议场景方面，通过优越的芯片性能构建一套基于SV100 芯片 的端到端方案，方案涵盖解码、图像处理、AI分析、视频拼接、编码等多个模块，为视频会议系统提供卓越的视频编解码、图像前后处理、声音处理能力，从而实现高清晰度、低延迟的会议体验，满足了现代商务沟通的需求；此外，在车路协同系统领域，基于瀚博SV100 芯片搭建的2D 视频 分析pipeline 和 3D 点 云 分析pipeline，实现了对路域环境的全量感知，提升了交通系统的智能化水平，为车辆导航、安全驾驶等方面提供了更为可靠的支持。



图43瀚博智能交通解决方案

资料来源：瀚博半导体

瀚博通过不断推出多样化的端到端解决方案，积极构建了一个庞大而完整的应用生态系统，为各行业提供了端到端方案参考，辅助客户基于瀚博半导体产品快速搭建其业务链路。

3. 4国内厂商在产品性能差距上正在缩小，但应用生态差距较大

3. 4. 1国产智算芯片在性能方面与英伟达的通用GPU 的差距正在缩小

虽然在绝对性能上，一些国产智算芯片可能还未完全达到英伟达顶级GPU 的水平，但差距正在逐步缩小，且在特定应用场景和性价比方面，国产芯片已经开始展现出独特的竞争优势。各厂商不断推出新的芯片产品，性能持续提升，每一代都在性能和能效比上有所突破。许多厂商采用专为AI任务设计的架构，而且通过与本土AI框架和软件生态系统的紧密合作，国产芯片能够在实际应用中发挥出更高的性能。随着国产智算芯片的不断发展，未来国产智算芯片有望在更多领域和场景中替代英伟达通用GPU，进一步缩小两者之间的性能差距。

表7主要智算产品性能对比

公司	型号	GPU架构	峰值INT8计算性能	峰值半精度(FP16)性能	显存容量	最大功耗	工艺制程	发布时间
英伟达	A100 80GB PCIe	Ampere	624 TOPS	312 TFLOPS	80GB HBM2	300W	7nm	2020 Q2
英伟达	H100 SXM	Hopper	3958 TOPS	1979 TFLOPS	80GB	700W	4nm	2022 Q1
AMD	MI100	CDNA	184.6 TOPS	184.6 TFLOPS	32GB HBM2	300W	7nm	2020Q4
谷歌	TPU v5p	TPU	918 TOPS	459 TFLOPS	95GB HBM			2023 Q4
海光	深算一号	GPGPU				350W		2021 H2
华为	昇腾310	达芬奇	16 TOPS	8 TOPS		8W	12nm	2018 Q4
华为	昇腾910	达芬奇	640 TOPS	320 TFLOPS		310W	7nm	2019 Q3
华为	昇腾910B	达芬奇		376 TFLOPS		400W	7nm	2023
寒武纪	MLU370-X4	Cambricon LUarch03	256 TOPS	96 TOPS	24GB LPDDR5	75W	7nm	2021Q4
壁仞科技	壁砺104P	壁立仞架构	1024 TOPS		32GB HBM2e	300W	7nm	2022 Q3
壁仞科技	BR 100	壁立仞架构	2048 TOPS	1024 TFLOPS	32GB HBM2e	300W	7nm	预计2023
天数智芯	天垓100	GPGPU	256 TOPS	128 TFLOPS	32GB DRAM HBM2	250W	7nm	2021Q3

资料来源：根据公开资料整理

3.4.2 国产智算芯片在应用生态方面还有很大发展空间

成熟的智算芯片厂商如英伟达、AMD 等不仅发展硬件产品，还构建了完善的生态系统，包括软件开发工具、应用程序接口(API)、开发者社区等。

在软件兼容和生态方面，占据主要市场份额的英伟达经过多年发展，英伟达的CUDA编程平台和库已经非常成熟。而国产智算芯片可能在软件生态方面还不够完善，导致开发者在使用时会遇到一些问题，需要一定的技术支持来完成模型的开发部署、最大化发挥芯片的性能。相比之下，国产智算芯片还需要更多的开发时间和资源来建立和优化自己的软件生态系统。

在开发者社区和资源方面，英伟达拥有庞大的开发者社区，提供了丰富的教程、文档、示例代码和论坛支持，这对于吸引和培养开发者至关重要。而国产智算芯片起步较晚，开发者社区相对较小，需要更多时间来积累经验和知识共享，引入更多的开发者和贡献者，不断完善软件生态，从而提高国产智算芯片的易用性。

在第三方应用支持方面，许多商业和开源的AI应用程序、工具和库都优先针对英伟达的GPU进行优化，这使得用户在使用这些应用时能够获得更好的性能和效率。国产智算芯片需要吸引更多第三方软件供应商和开源项目为其提供支持和优化。

在市场认可度方面，英伟达等国外厂商的品牌影响力较高，加之开发者对CUDA 生态的熟悉度和接受度，使得很多情况下将英伟达作为首选产品。国产智算芯片需要通过持续的技术创新、优质的产品和服务，以及成功的应用案例来提升自身的品牌形象和市场接受度。

四、智算产业发展趋势

4.1 多元异构算力融合与国产化是方向与趋势

4.1.1 智算芯片是智算产业的核心环节

通用算力、超算算力与智算算力三种算力适用的应用场景有所不同。通用算力主要来源于通用处理器CPU，应用场景非常广泛，可用于对精度要求不高的数字化场景中。而超算算力则主要支持天体物理、气象研究、航空航天等高精尖科研领域。智算算力则主要应用于AI场景中。从精度看，可把算力分为双精度(FP64)、单精度(FP32/TF32)、半精度(FP16/BF16)及整型精度(INT8/INT4)。对于AI模型训练及推理来说，处理文字、语音、图片或视频等需求较大多，以单精度、半精度、整型计算精度为主，少量储备双精度能力，即可满足绝大部分应用需要。一般来说，相比于模型训练，模型推理所需的算力精度较低，很多场景FP16、BF16、INT8即可满足需要。

产业蓬勃发展，呈现出对多元异构算力融合的需求。异构算力是指包括CPU、GPU、FPGA、ASIC等多种算力在内的处理体系，能够协同工作以满足不同场景中的应用需求，实现计算效力的最大化。在 market 需求的推动下，算力的发展呈现出多样性的趋势，打破了传统的单一架构，使异构算力能够更好地适应不同场景下的数据处理应用。此外，异构算力的发展也表现出能力开放和统一管理的趋势。无论是芯片厂商还是平台厂商，它们都围绕自身的产品系统，将底层的异构算力能力进行融合，以吸引更多的产业链上下游企业参与，共同打造更为健全的生态环境。目前，异构AI算力发展局限，面临异构芯片适配标准尚需统一、异构硬件移植适配亟待优化、软件生态技术应用相对薄弱等问题。

多元异构算力的整合要求在“横向”和“纵向”两个方向上实现协同发展。在“横向”方面，通用算力、智能算力以及超算算力正在向多元异构的方向发展，满足多样应用需求。可将CPU、GPU、FPGA、ASIC等多种芯片实现有效组合，以达到通用算力与专用算力的有机融合。目前许多应用需要不同的算力，有时需要通用算力，有时需要智算算力，有时又要用超算算力。如何合理、精确地计算客户应用所需的算力服务或资源成为关键。如果计算场景与算力精度之间产生错配，使用高精度算力计算低精度需求场景则会功耗过高，造成资源浪费，反之则导致计算任务无法完成。因此，加强算力精度与计算场景匹配，推动多元异构算力融合对提高算力资源利用效率、国家一体化算力网络建设具有重要意义。相较于垂直发展专用智能算力的路径，全面提升多元融合算力水平，通过混合精度算力输出满足数据模拟、模型训练、智能推理等多样需求更能满足现代发展需要。同时，各地各主体在算力中心建设中，也注重构建融合通用计算、智能计算和超级计算的平台，以促进多元算力资源有机协同，为用户提供平台级服务。如中国电信天翼云可提供智算、超算、通算多样化算力服务的一体化算力平台，赋能客户使用算力就像“水电”一样无感普惠，同时还能完全满足客户需求。中科曙光在2020年正式上线全国一体化算力服务平台(ac.sugon.com)，实现了对全国十多个算力中心资源的整合与统一任务调度，聚合不同精度算力。

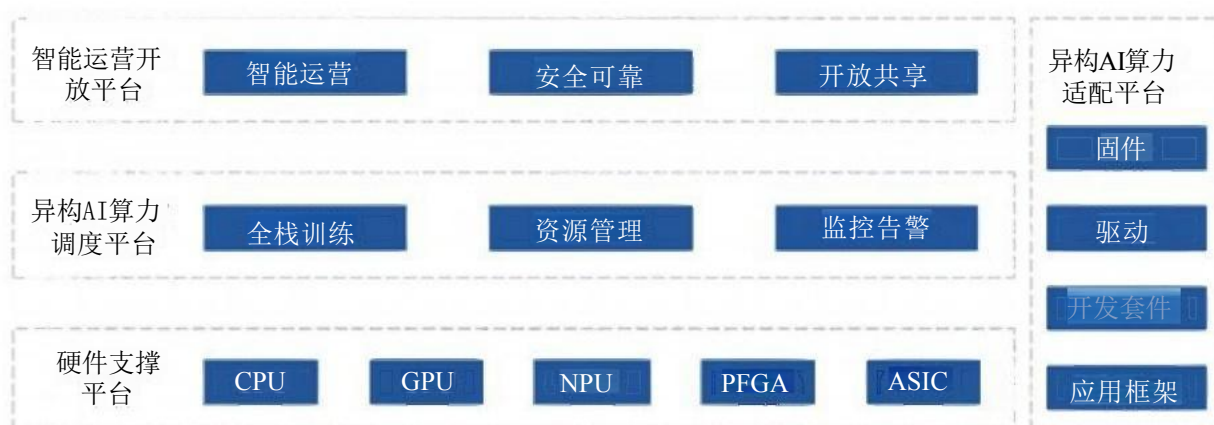


图44 异构AI算力操作平台架构

数据来源：中国信通院郭亮等：异构AI算力操作平台的架构设计与优化策略

在“纵向”方面，着眼于开放与协同，通过采用开放架构实现对成熟主流软件生态的兼容，支持主流人工智能框架、算法模型、数据处理技术及广泛的行业应用。如中科曙光的AI基础软件系统DAS，从AI组件的纳管、性能组件的优化、软件工程化能力的建设等多方面着手，聚焦打造一套完整的人工智能软硬件生态体系，帮助用户解决模型训练效率低下、算法部署困难、算力资源利用率不高等问题。目前曙光DAS已全面适配了市面上不同的框架、加速库等，并开发了快速适配工具，支持软件组件高速迭代。这将极大提高平台的适用性，增强“基础设施”的易用性和通用性，推动整个AI产业链的多元化、协同化和繁荣发展。

中科曙光人工智能基础软件系统



图45 曙光人工智能基础软件系统DAS

来源：中科曙光

2022年，美国商务部宣布实施AI芯片出口管制，英伟达推出阉割版芯片H800与A800。2022年10月，美国政府宣布实施人工智能芯片出口限制，同时满足芯片的I/O带宽传输速率大于或等于600 Gbyte/s、“数字处理单元原始计算单元”每次操作的比特长度乘以TOPS计算出的算力之和大于或等于

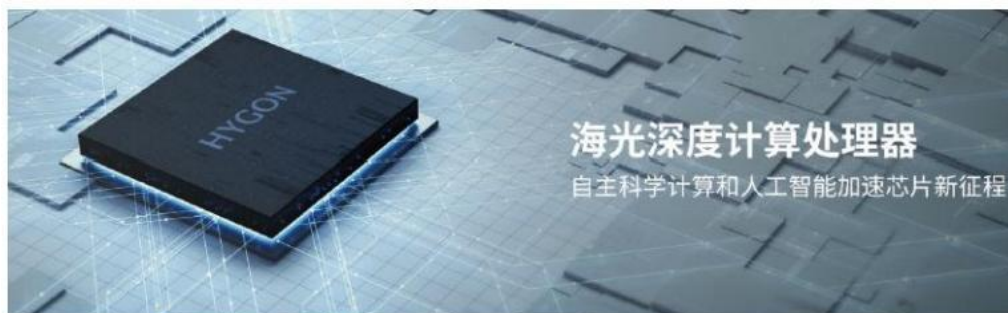
4800TOPS 两个条件的芯片为受管制的高性能计算芯片。这也使得英伟达无法向中国市场提供A100和H100 等高端AI芯片，随后英伟达在遵守美国限制规则的前提下降低带宽传输速率，向中国出售特供版A800 和H800。2023年10月，美国政府发布一系列新版芯片出口限制，并取消30天豁免期，立即限制英伟达向中国出口H800、A800, 相关芯片需求无法得到满足，价格增长较多。同时，我国AI芯片不断迭代发展，性能不断提升，有望持续获益。另外，在英伟达H800、A800等芯片的供货受限的背景下，英伟达的GPU 卡与众多国产AI加速卡组集群，进行异构算力融合也成为发展趋势之一。政府不断出台相关政策推动智算产业国产化发展。在智算算力国产化方面，我们着重加速人工智能公共算力平台的建设，致力于提供可靠高效的本土算力。截至2022年6月，科技部已经印发《国家新一代人工智能公共算力开放创新平台建设指引(试行)》(以下简称《指引》), 为各地在建设公共算力平台方面提供了明确的指导。《指引》中对平台的国产化程度、算力规模与解决方案等方面进行了明确要求。要求自主研发芯片提供的算力标称值占比不低于60%, 并鼓励优先采用国产开发框架，使用率不低于60%。此外，在算力性能方面，《指引》对AI训练和推理的常用规格提出了具体要求，包括16位浮点 (FP16) 性能应达到400PFLOPS,32 位浮点 (FP32) 性能应达到200PFLOPS,16位整型 (INT16)性能应达到400POPS。与此同时，《指引》针对当今大规模模型训练需求激增的情况，对环境承载能力也提出了具体要求，强调平台应配置成熟易用的人工智能全栈运行环境，以确保能够运行千亿级参数的预训练模型。

华为、海光信息、寒武纪等巨头坚持迭代升级，其产品性能日益提升，有望在AI训练算力领域获得更多市场份额，实现国产替代。另外，相比于AI训练芯片，推理芯片的性能要求以及精度要求较低，部分国产AI芯片凭借其良好性能以及性价比能够满足推理端的需求，我国AI芯片国产化比例有望提升。目前，华为新一代AI芯片昇腾910B发布，芯片性能达到英伟达A100 芯片水平，打破美国在人工智能芯片领域的技术垄断地位。海光信息新一代DCU(GPGPU) 深算二号已经发布实现商用，可用于大模型训练和推理。寒武纪新品思元590也得到性能大幅提升。



· 图45华为AI处理器昇腾910及310

资料来源：澎湃新闻



产品优势

强大算力

基于主流通用并行计算架构，可招配海光CPU使用

广泛应用于科学计算，人工智能模型训练和推理

开源生态

· 开放式生态，统一底层硬件驱动平台，支持常见计算框架、库和编程模型

· 层次化软件栈，适配不同API接口和编译器，可最大限度利用已有的成熟AI算法和框架

图46海光AI芯片性能强大、生态优异

资料来源：海光信息

4.2 智算建设与东数西算工程将有效融合

2022年2月，作为算网建设开始的“东数西算”正式启动。2022年2月17日，国家发改委、中央网信办、工业和信息化部、国家能源局联合印发通知，同意在京津冀、长三角、粤港澳大湾区、成渝、内蒙古等8地启动建设国家算力枢纽节点，并规划了10个国家数据中心集群。至此，全国一体化大数据中心体系完成总体布局设计，“东数西算”工程正式全面启动。其中，八大枢纽由五大西部枢纽以及三大东部枢纽组成：内蒙古、宁夏、甘肃、成渝、贵州、京津冀、长三角以及粤港澳枢纽。十大集群由六个西部集群以及四个东部集群组成：和林格尔、中卫、庆阳、天府、重庆、贵安、张家口、芜湖、长三角生态绿色一体化发展示范区、韶关。

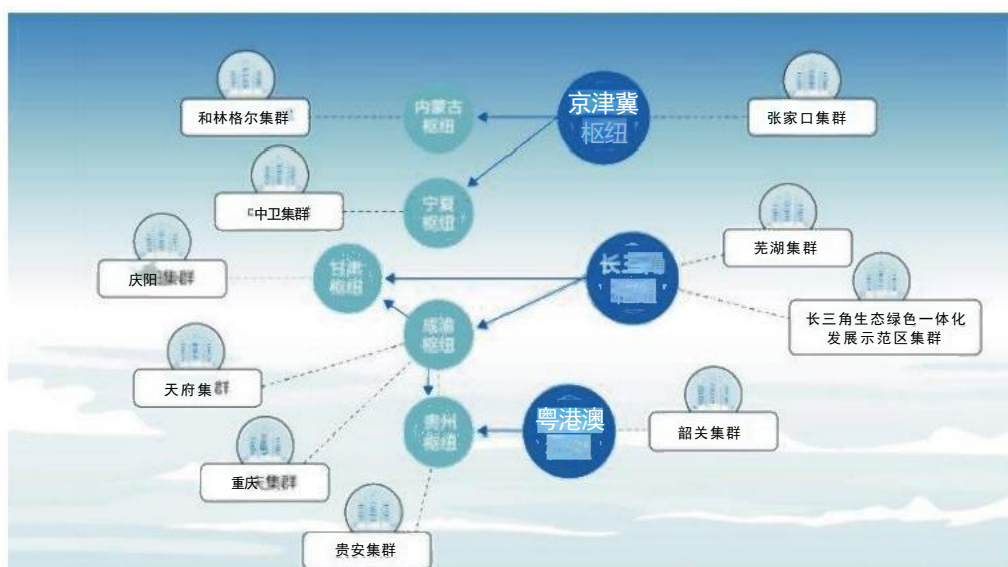
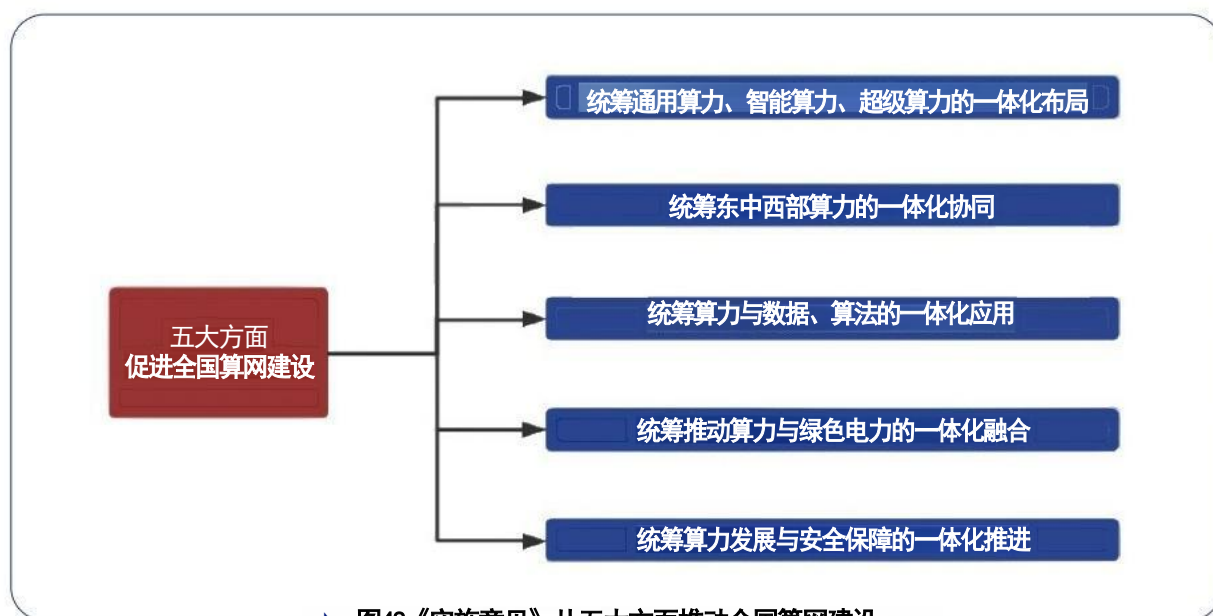


图472022年2月，东数西算正式启动，8地建设国家算力枢纽节点，规划10个国家数据中心集群

资料来源：华为

东数西算工程还需要进一步深化发展。首先由于缺乏统筹规划，各地政府在智算中心建设方面存在的无序扩张问题，需要统筹通用算力、智算、超算的一体化发展，以解决盲目扩张、资源浪费等风险。其次，西部数据中心空置率较高，需要完善全国一体化算力网的布局，以更彻底、完善地解决西部数据中心的空置情况。第三，目前东中西协同机制存在不足，东部枢纽节点应更积极参与和主动建设，推动东西之间的协同共享。最后，西部枢纽节点地区绿色电力使用率较低，需要提高西部地区可再生能源利用率。

2023年12月29日，国家发改委、国家数据局、中央网信办、工业和信息化部以及国家能源局正式发布《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》（以下简称《实施意见》）的全文。《实施意见》从五方面对全国一体化算力网建设提出了指导意见：强调了统筹通用算力、智能算力、超级算力协同计算；东中西地区及大中小城市协同布局；算力、数据、算法协同应用；算力和绿色电力协同建设；算力发展和安全协同保障。同时，《实施意见》设立了到2025年底综合算力基础设施体系初步成型、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、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等目标。“东数西算”、全国一体化算力网络建设再深化。



→ 图48 《实施意见》从五大方面推动全国算力网建设 ←

资料来源：发改委等《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》、东方证券研究所绘制

此次发布的《实施意见》强调了国家枢纽节点外原则上不得新建各类大型或超大型数据中心，进一步推动各类新增算力向国家枢纽节点集聚，将国家枢纽节点打造成为国家算力高地。同时，《实施意见》提出了到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的建设目标。同时，国家持续推动统筹东中西部算力的一体化协同，提升国家算力网络的传输效能，着眼于推进枢纽节点内、节点间以及枢纽节点与非枢纽节点之间高度确定性与高通量网络的构建。《实施意见》中确立了到2025年底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等目标。确定性网络在多地多领域展开实践性探索，各地正针对全新的网络架构对未来网络实验设施(CENI)进行成果验证。如山东已建成并运行覆盖16个地市、5600公里的全球首张确定性网络，其核心节点时延抖动控制在20微秒以内，主要性能指标达到国际领先水平。

4.3 通过算力调度，智算由单个节点向多点融合与协同演变

国家持续鼓励探索算力调度运营的机制，在《实施意见》中强调建立跨区域算力资源调度的机制，探索开展城市算力网建设，实现枢纽节点算力资源与城市算力需求的高效供需匹配，打造跨区域的算力调度体系。算力网推进路径涵盖了从城市到区域再到全国的层级，城市算力网的建设被视为促使省域或区域性算力网横向延伸的关键动力。城市算力网可推动本地闲置算力资源加入网络，令非枢纽节点的地方与枢纽节点可以建立联系，为算力提供了横向连接的契机，并加强了国家级算力枢纽节点的储备功能。同时，国家促进充分利用枢纽节点的算力资源联动调度枢纽节点、节点外中西部绿色能源丰富地区的算力资源，鼓励以点对点结对子的方式来推动东西部算力的协同调度。

算力调度是指通过对不同业务的算力资源和算力需求进行匹配，使合理的算力去处理相应数据的一种方式。理想情况下，算力调度可以实现跨服务商、跨架构、跨地域的算力任务部署和结果交付，可实时感知客户的算力需求，并根据算力资源池的情况，给客户自动调配符合其需求的最优算力，获取算力将像获取水、电一样便捷、价低。算力调度对提高计算机资源的利用率，减少资源浪费，保证任务的高效执行有着深远的意义。

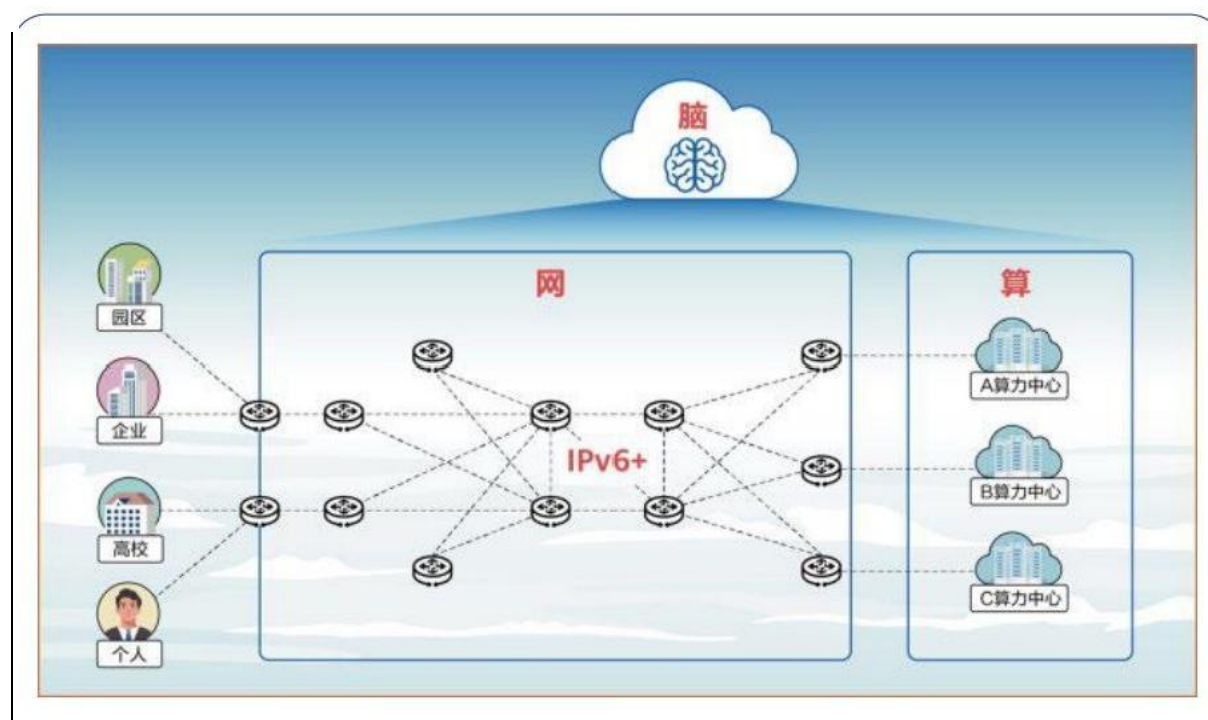


图49算力调度是算网组成的关键

资料来源：华为

根据于美泽等的《算力调度关键问题和实施路径研究》，从算力调度的实现流程来看，可将算力调度的关键技术划分如下：

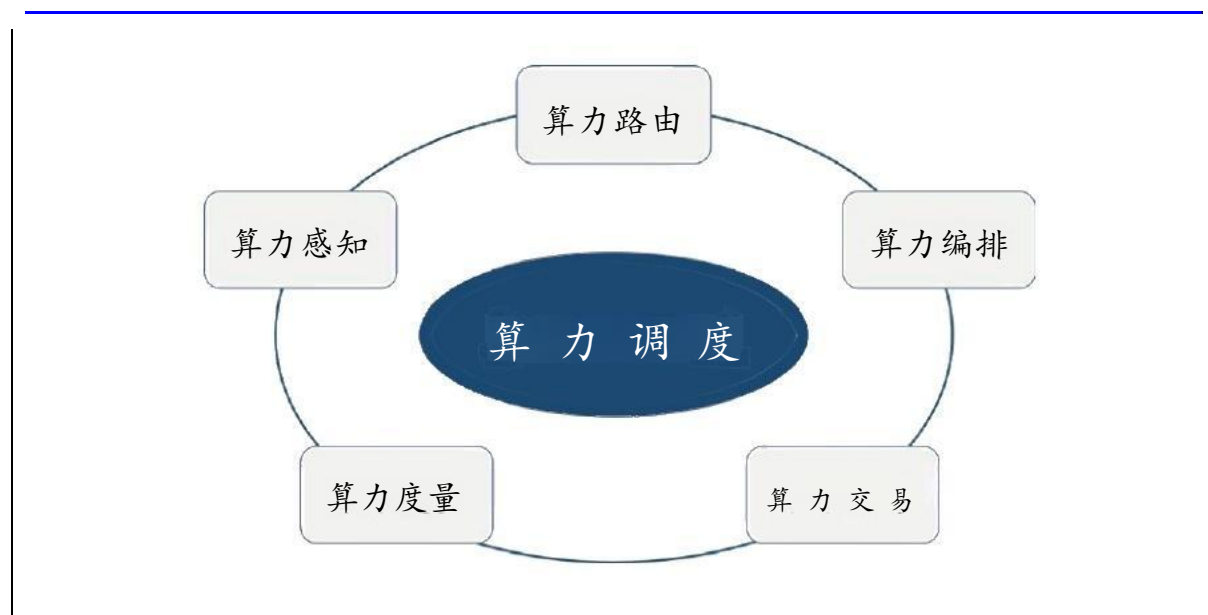
●**算力感知**，是指对算力提供方的计算、存储、通信、服务等不同类型资源与服务，以及对算力需求方业务需求信息的感知。一方面，分布式算力节点将通过网络，将感知到的算力特性的处理能力、负载能力、部署位置等要素上报至算力感知中心；同时，算力感知中心也将主动查询算力节点的各类性能。

●**算力度量**，是指建立统一算力标准对算力进行量化。算力的复杂性导致难以通过单一维度来量化，并且其复杂性主要体现在三个方面。首先，CPU、GPU、FPGA、ASIC 等多种计算单元组成的异构处理体系难以进行标准化的统一；其次，网络、存储等多维资源，本身就需要从多个维度进行建模和评估；最后，不同行业、不同企业、不同场景对算力资源类型和需求量的要求存在较大差异，用户很难精准把握需要多少算力。作为灵活调度多元异构算力资源的前提，从建立算力资源模型出发，构建算力度量标准势在必行。

●**算力路由**，是指整合算力资源信息、构建路由表以动态匹配和连接调度算力任务，以最优方案解决具体算力调度问题的技术。在算力路由方面，目前主要采用的集中式路由部署策略在实时性方面表现欠佳，分布式路由协议下的计算优先网络将有望取而代之。

●**算力编排**，该技术基于算、网、数的原子能力的按需组合完成算力业务的路径编排，并通过网络控制器进行算力路由选择与转发，实现算网业务的统一编排、部署和保障。算王编排是算网大脑的核心控制部分，具备云原生编排能力、多量纲的编排能力以及运行态的编排能力。

●**算力交易**，是指算力提供方与算力需求方通过算力交易平台进行资源交易的新型商业模式。算力交易平台基于“身份、协议、订单、账单、佣金”的“五信”交易体系，根据用户的差异化需求，实现智能、公平、泛在、可溯、可信的算力交易。



→ 图50算力调度关键技术

资料来源：于美泽等. 算力调度关键问题和实施路径研究[J]. 信息通信技术与政策, 2023, 49 (5) :9-14、东方证券研究所绘制

算力调度对提高计算机资源的利用率，减少资源浪费，保证任务的高效执行有着深远的意义。一方面，算力调度为开展算力相关活动提供了基础支撑。以集群模型训练为例，由于训练模型的巨大算力开销，本地单一的智算中心可能无法满足模型训练需求，通过算力调度，该问题将能得到有效解决。另一方面，算力调度能够实现算力需求和供给的精准匹配。由于算力需求与供给的地域不平衡性，存在部分地区算力供给过剩，同时部分地区算力需求无法得到有效满足的现状。通过算力调度，全局范围内的算力供给与需求有望达到平衡，算力资源的利用率将得到显著提高。建立健全算力调度网络、推动智算由单节点向多节点融合和协同演变具有重要意义。

目前，算力调度面临着跨服务商、跨架构、跨地域的三大挑战。算力调度目前面临着诸多问题，例如服务业务迁移与数据流通有壁垒、跨地域的算力资源调度门槛高、跨架构的算力资源协作不足等，同时各地自行建设算力调度中心，整合难度大调度难统一。

算力网络已成为国家重点发展领域之一，算力调度更是其中关键。在《实施意见》出台前，科技部于2023年4月17日启动了国家超算互联网工作，成立了国家超算互联网联合体，并提出预计到2025年，国家超算互联网将形成总体布局，成为数字中国建设的“高速路”。目前，我国有十余个国家超算中心，但是目前基本处于“孤岛”状态。国家推动打造超算互联网成为淘宝、京东一样的平台，让各行各业的超算应用APP如雨后春笋般发展，使用户能够方便、高效地使用超算。

宁夏、南京、上海等多地政府先试先行，打造算力调度平台、枢纽城市。2023年2月，宁夏上线了国内首个可以支持算力交易调度的应用系统“东数西算一体化算力服务平台”，能为智算、超算、通用算力提供算力发现、供需撮合、交易购买、调度使用等服务。中科曙光凭借自己在算力调度运营方便积累的深厚技术和实践经验，在整体方案设计、平台建设、算力调度以及统一算力接口等方面为该平台提供了重要技术支持。2023年2月，南京城市算力网运营平台正式发布，实现了联邦模式的“算力并网”、“全局调度”、“异构算力统一度量”等关键技术，为国家“东数西算”工程重大战略赋能助力。2023年2月，上海市人工智能公共算力服务平台正式揭牌投用，将更好满足科研机构 and 广大中小微企业实际算力需求。2023年4月，《郑州庆阳哈密协同建设城市算力网实验场合作协议》正式签署，三方将聚焦算力网建设关键，联动打造智能算力新引擎，推动构建以庆阳为“算力供给站”、以郑州为“算力中继站”的数算电融合“东数西算”工程。2023年5月，贵州省发布算力运营调度平台2.0版，从数网、数纽、数链、数脑、数盾五个方面，构建起了覆盖能力、质量、结构、通道、产业的指标体系。2023年7月，北京经济技术开发区推出了我国首个“算力资源+运营服务+场景应用”一体化建设工程“公共智能算力中心启动和算力调度服务平台”，前者总体规划1000P建设规模，以“微利、惠企”原则构建新型算力基础设施，后者则是国内首个超大规模、多源异构、多算一体、安全可信、生态融合的算力调度服务平台。



●图51宁夏发布东数西算一体化算力服务平台、贵州枢纽算力调度平台

资料来源：新华网、贵州省人民政府

中科曙光、中科院计算所、鹏程实验室、华为等诸多主体单位积极投入算力调度平台开发，探索算力调度运营模式。中科曙光是多源异构算力跨域调度与一体化运营服务模式的倡导者，早在2020年就上线了全国一体化算力服务平台(ac. sugon. com)。2022年10月，中科曙光推出全国一体化算力服务平台4. 0, 对平台设计、核心功能、交互方式和后台技术等进行了重构，并新增了跨域传输、快捷入口等多个功能，以“更直观”，“更自由”，“更高效”的性能表现，通过输出算力-存储-数据-环境-应用为一体的服务，加速海量复杂行业应用创新与落地。

共享资源

异构节点-4D2 华东三区【乌细】西北-X【西安】	7285-32C-128G-4卡-3 编号	DCU 类型	1-728532C 2.0GHz 处理器	28GB 内存	200Gb 计算网络
异构节点-4D1-2 东北一区【确】	7285-32C-128G-4卡-2 编号	DCU	1*728532C 2.0GHz 理	28GB	200Gb 计 网络
异构节点-4D1 华旅-区【R山】	7185-32C-128G-4卡	DCU 类型	1*718532C 2.0GHz 理	28GB	200Gb 计算网络
异构节点-4R1 华北-区【面】	7185-64C-256G-4*BIX3080 编号	GPU	2*718532C 2.0GHz 处理器	56GB 内存	00G 计算

• 图52曙光智算提供多地异构算力

资料来源：曙光智算

中科院计算所方面，信息高铁建设如火如荼，为算力流通打下了良好基础。自2019年以来，在孙凝晖院士的带领下，围绕信息高铁，计算所建立起包含卫星网络、工业5G 网络、面向管理的全局测调中心以及面向用户的全局运营中心的信息高铁综合试验场的内场，以及全国多地建设的包含超算中心、智能计算中心、高通量中心、大数据中心等在内的不同类别的信息高铁站。各地的企业、政府可通过计算所所接入到算力网络中，获取所需算力及服务。信息高铁的建设可以真正使算力流通起来，打破垄断，普惠中小企业及大众。

鹏城实验室方面，其所打造的“中国算力网(China Computing NET, C²NET)”，已接入全国多个超算、智算节点。2022年6月15日，鹏城实验室联合华为共同打造的“中国算力网-智算网络一期”正式上线，目前已有鹏城云脑、西安、武汉、青岛等地20多个包括人工智能计算中心、超算中心、一体化大数据中心在内的节点接入中国算力网。华为方面，所发布跨区域算力调度IP网络解决方案，智能云图算法可将算力运输效率提高30%以上。华为提出的智能云图Edge-Disjoint KSP算法，基于云因子与网因子一体化调度，可根据企业的需求选择最佳云池，同时实现多源到多宿的云网资源平衡，将算力输送效率提升了30%以上，有效解决了传统广域网采用的最短路径调度造成的链路利用率不均衡、多路负载分担的业务对网络的差异化需求问题以及过往未考虑的云池因素等问题。

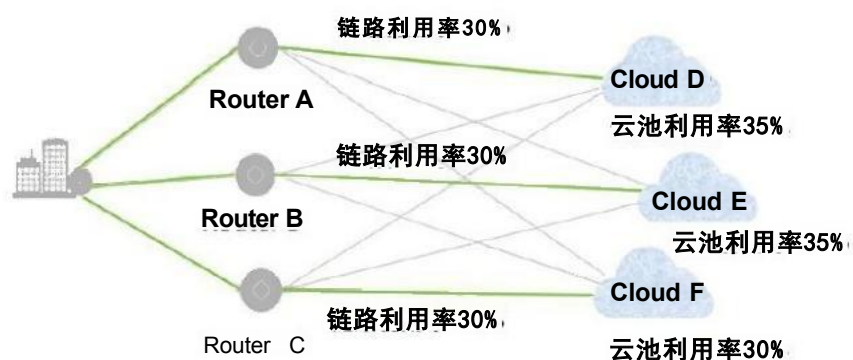


图53广域网络的云网一体化调度

资料来源：华为

此外，中国信通院联合中国电信共同发布了首个实现多元异构算力调度全国性平台“全国一体化算力算网调度平台(1.0版)”、并行科技打造算力网络资源池上线三大运营商算力资源、中软国际打造了以“算力接入管理中心”、“算力编排调度中心”和“算力运营中心”为核心内容的中软国际算力调度平台、青云科技发布青云A1算力调度平台与青云A1算力云、首都在线基于K3s 自研算力调度系统AiDN, 发布了“新一代算力平台”。众多企事业单位在算力调度领域均有积极行动。



图54中国信通院联合中国电信共同发布的全国一体化算力算网调度平台

——智算应用场景展示

资料来源：中国工信新闻网

4.4 智算中心低碳化发展需要算力-电力协同

2021年11月，国务院印发《2030年前碳达峰行动方案》，指出要重点实施能源绿色低碳转型行动、节能降碳增效行动等“碳达峰十大行动”。而数据中心已经成为电力的消耗大户，而随着人工智能的发展，智算的需求增长会持续拉动对电力的需求增长。据《中国“新基建”发展研究报告》，到2025年，预计未来数据中心将占到全球能耗的最大份额33%。在“双碳”目标下，国家对数据中心集群作为大型电力工业用户和新负荷主体的电能使用效率和可再生能源利用率有更为严格的要求。如何便宜、稳定、持续地使用绿电成为发展的关键之一。推动算力、电力统筹规划，东数西算、东电西送有机协同，是实现双碳目标的重要路径。

目前，电是制约算力设施发展的主要矛盾，算力-电力协同的重要性日益显现。电力与算力的影响是双向的，一方面，算力电力时空方面的错配、持续高涨的成本压力、ESG/双碳管理的压力，制约了算力设施的发展。同时，随着智能算力的暴增、算力侧很少考虑电网的安全，电力电网的稳定运行也面临着挑战。在此背景下，深入推进算力协同，将成为建设国家算力网，推进智算中心建设的关键一环。

算力-电力协同，即数据中心与电力系统协同，是指通过数据中心的灵活用能，来赋能电力系统源网荷深度互动，通过电力市场信号引导数据中心用户灵活调度算力负载执行时间、地点，建立起原动力、调配力、平衡力、盈收力、赋能力的“五力协同”机制，让算力端享受更便宜、绿色、及时、可靠的电，同时让电力端安全高效、可预测、可管理、柔性灵活、可调度，以电力与算力的高效协同与融合发展，助力提升电力系统运行的灵活能力，降低数据中心用能成本和碳排放。

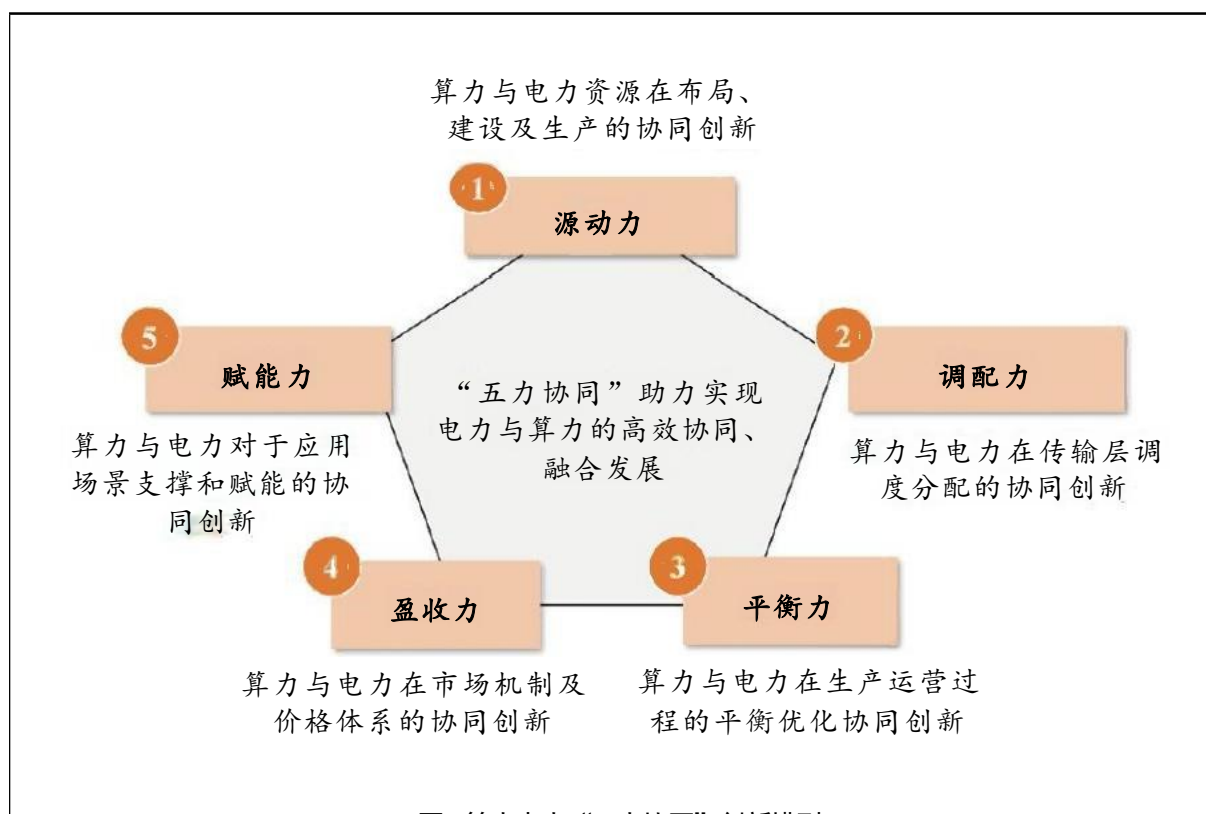


图55 算力电力“五力协同”创新模型

资料：侯金刚，算电协同的重点和突破

算电协同可划分为三大层次，包括算力节点与区域电网协同、算力网与电力网协同、算力市场与电力市场协同。算力节点与区域电网协同层面，可通过技术框架协同，打造新一代算力中心或园区。通过园区储能，支持动态负荷管理与反向调度。同时加大绿电直供与绿电交易比例。第二是算力网与电力网协同，以实现双网协同调度与多时空维度调度。第三是算力市场与电力市场两个万亿级市场的协同，以实现算力价格与电力价格的实时动态协同，以及各类算力服务产品与各类电力产品现货、期货的协同。

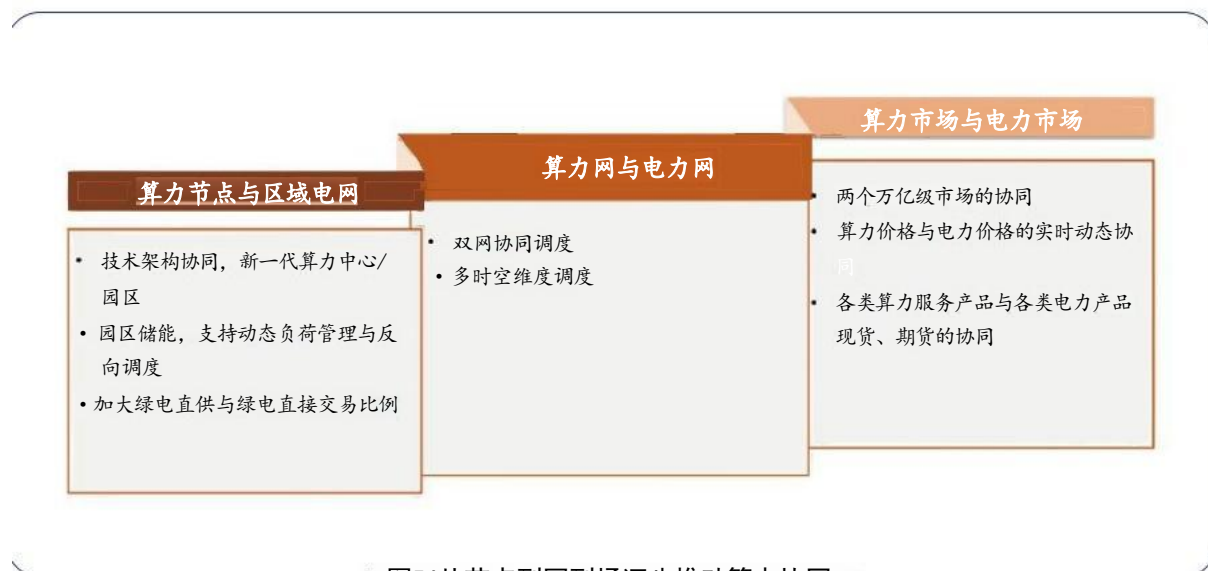


图56从节点到网到场逐步推动算电协同

资料来源：粤港澳大湾区大数据研究院

推进算电协同，一方面有助于数据中心节能降耗，同时通过接入绿色电力，将有效解决当前绿色电力使用率低的问题。统筹推动算力与绿色电力的一体化融合，开展数据中心智慧能源管理，提升废弃物绿色回收与循环再利用的水平，创新算力电力协同机制以促进绿色电力与算力的高效对接，将会是建立全国一体化算力网络的必不可少的举措。

2023年12月，国务院印发《实施意见》，就算力与电力的协同问题做出了重要指示。针对目前智算中心无法直接使用成本低廉的绿电的问题，《实施意见》鼓励在枢纽节点地区，通过源、网、荷、储一体化的方式促进绿电直供，同时强调要积极推进数据中心的智慧能源管理，推动数据中心储能设施反向作为送电、发电的单元，充分释放储能的算力。此外，数据中心的循环利用，即服务器、机电设施等更新迭代后相关零部件的循环利用，也有待探索。

河南、甘肃等地方也就算电一体化协同发展先行先试。2023年7月28日，河南省发布《河南省人民政府办公厅关于印发河南省支持重大新型基础设施建设若干政策的通知》，强调鼓励超算中心、智算中心、新型数据中心等因地制宜地配置储能设施，以提升电力自平衡能力，有力推动了算力基础设施与储能设施的融合发展。此外，文件指出，对配套建设储能规模在1000千瓦时以上的储能设施，省财政将给予一次性奖励。此举有望对企业发展算力基础设施产生显著激励作用，进而推动储能技术的应用。

2023年12月，在全国一体化算力网络国家枢纽节点（甘肃）庆阳数据中心集群绿色普惠算力行动推进大

会上，甘肃省副省长陈得信强调，要统筹甘肃所能，做好数算电融合文章，通力打响甘肃省绿色能源与算力融合发展招牌。会上，国家发展改革委价格监测中心副主任王建冬作了《通用人工智能背景下创新数算电价联动机制的初步思考》的演讲，建议庆阳市主动探索数算点融合发展新路径，创新电力和算力价格联动结算方式和促消纳来降低成本，建立起算力+算法+数据综合定价机制为主攻方向，打造全国一体化算力建设中的“庆阳模式”，为“东数西算”工程做出更大贡献。

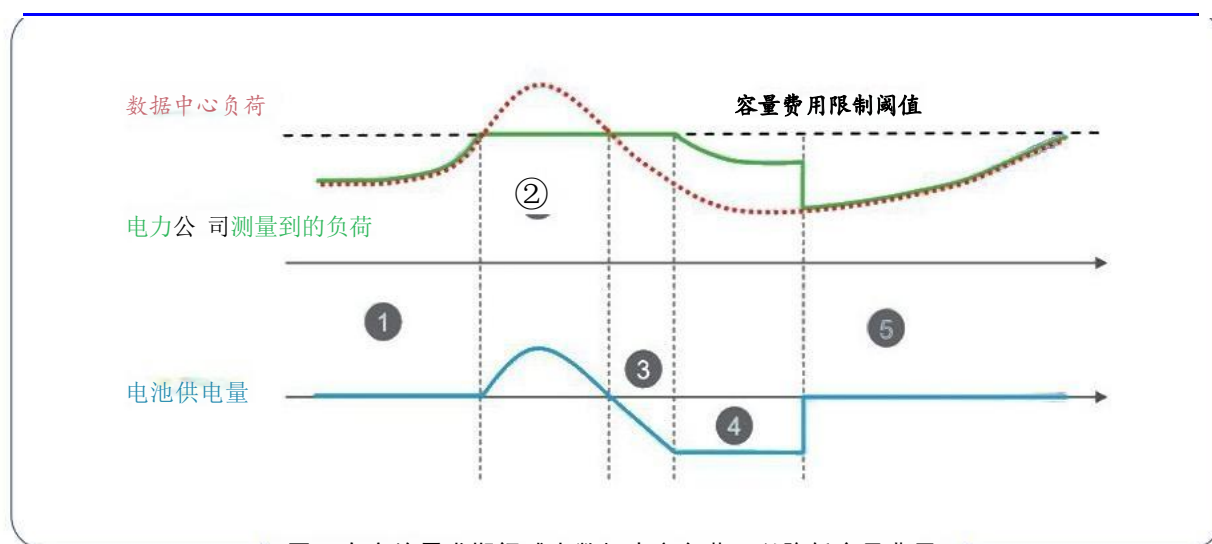


图57在高峰需求期间减少数据中心负荷可以降低容量费用

资料来源：施耐德

未来，可进一步同东数西算各枢纽节点充分调研、沟通算电协同方面的难点、堵点和兴趣点，借助国家算力网深入实施的政策文件，在枢纽节点选择存量和增量的数据中心园区实施算电协同的方案和机制，同时充分利用算电协同创新中心的体制机制优势，同当地政府、算力园区业主/客户和电网广泛合作，以研究促工程，以试验推政策创新。相信在中央《实施意见》的顶层设计下，在各地政府奋勇争先、先试先行的积极探索下，可以预见我国数电融合将取得长足发展，以更低、更清洁的能源消耗助力双碳目标的实现。



图58算力电力的双向协同、高效融合是算电融合的目标

资料来源：施耐德

五、优秀案例介绍

0=

5.1 中科曙光

中科曙光作为我国核心信息基础设施领军企业，为中国及全球用户提供创新、高效、可靠的IT产品、解决方案及服务，公司于2014年在上海证券交易所上市(股票代码：603019)。经历20余年发展，中科曙光在高端计算、存储、安全、数据中心等领域拥有深厚的技术积淀和领先的市场份额，并充分发挥高端计算优势，布局云计算、智能计算、绿色计算等领域的技术研发，打造计算产业生态，覆盖“芯、端、云、算”全产业链，为科研探索创新、行业信息化建设、产业转型升级、数字经济发展提供了坚实可信的支撑。

打造完整的IT基础架构产品线

经历多年发展，曙光公司在高端计算、存储、安全、数据中心等领域拥有深厚的技术沉淀和领先的计算优势，拥有完整的IT基础架构产品线。

其中，高端计算机产品主要包括机架式服务器、高密度服务器、刀片服务器、超融合一体机产品等，能够面向多种应用场景，兼顾性能、能效、应用生态，具有领先的计算密度和节能性，产品整合高速网络和存储技术，可实现超大规模线性扩展。公司的高端服务器产品也涵盖浸没液冷、冷板液冷等产品形态，具有节能高效、安全稳定、高度集成等特点。随着国产化替代政策的深入，公司推出的国产高端计算机产品，已广泛应用于运营商、金融、能源、互联网、教育等行业客户，覆盖基础设施、电子政务、企业信息化和城市信息化等领域。

曙光存储产品包括分布式统一存储、多控统一存储、备份一体机等。公司分布式统一存储系统ParaS-tor历经近二十年自主研发，涵盖海量数据管理、高速网络、存储协议栈优化等核心技术，具有数百PB级超大规模部署实践经验；全闪分布式存储可以为人工智能、科学计算等场景提供极致性能及成熟的数据服务。近两年，公司存储产品线在AI大模型存储、液冷存储研发等领域实现突破。AI大模型存储解决方案提供多级缓存加速、XDS 数据加速及智能高速路由等多种数据IO 性能优化能力；公司推出业界首款液冷存储产品，实现冷板式液冷与存储技术深度结合，存储系统PUE 值降至1.2以下。2023年，公司以两个标段第一中标中国移动分布式存储集采，同时中标中国移动新型智算中心、华南理工大学液冷存储等项目，为国内多个行业AI大模型提供了极致性能的全闪存储平台，并携手中国移动成功打造行业首个存力智能调度平台。

曙光网络安全产品包括流量分析、网络安全平台、工业控制等国产软硬件产品及解决方案。公司研制并迭代以汇聚分流、网络流量分析系统、网络与计算加速以及行业应用系统为主的多种流量分析系列产品，为多行业客户提供网络流量可视化、流量回溯分析、5G信令解析等软硬件解决方案。目前，流量分析产品已广泛应用于政法、运营商、金融、科教与其他企业级客户。国产化安全平台已与多个头部安全厂商达成战略合作，共同打造高性能国产化安全产品，并面向数据中心防护、等保、企业安全、网络审计等多场景推出产品与方案。2023年，公司发布了工业数智平台“曙睿 SugonRITW”，并以此为基础，提出“以数智底座，加速数实融合进程，赋能新型工业化”的智慧工业战略。

曙光自2007年开始从事云计算技术产品研发，具有长期的技术积累，目前已形成全栈云服务能力。Cloudview云计算操作系统和StackCube-K 超融合系统等多款产品获得可信云首批认证。曙光云以建设运营“城市云”为主要业务模式，在全国多个城市布局，基本形成了安全可信的城市云服务体系，基于城市云计算中心为政府和企业用户提供云服务及云技术服务。

积极推进5A级智算中心建设

近年来，面对智算中心等大规模建设需求，曙光深度参与数字基础设施建设，同时不断提升智算中心解决方案综合能力，基于浸没液冷、高密度设计、高效电源管理、大规模算力调度平台、自动化运维、清洁能源等诸多先进技术，为用户提供整体智算基础设施解决方案，支撑智算中心集约化、规模化、绿色化高质量发展，赋能多区域多行业数智化发展。



● 图：曙光在2021世界人工智能大会上发布新一代人工智能计算平台

中科曙光5A级智算中心以其“开放、融合、绿色、普惠、服务”的鲜明特质，构筑起覆盖算力供给、算法优化、数据服务及行业应用的全场景智能计算服务体系，引领AI技术与各行各业实现由点至面的深度融合，为需求侧带来前所未有的易用性、通用性、经济性、节能性和省心体验。

5A级智算中心的核心竞争力，源自中科曙光自研国产异构智能计算硬件资源。硬件平台不仅能无缝应对数值模拟、AI训练、AI推理等全精度、全场景计算任务，确保计算效能与安全性并举，更以其自主可控特性，为国家信息安全筑起坚实屏障。公司承建的多个智算中心采用众多前沿技术：智算中心解决方案内的存储系统采用全自研ParaStor 分布式全闪存储，以自研技术支持AI芯片高速数据直接存取，优化存

算协调，将整体I/O性能提升数倍，满足智能计算性能需求；解决方案内的液冷系统，基于公司多年的技术积累，在计算核心部件实现液冷散热时，首次实现液冷存储散热系统应用，降低能耗的同时大幅提升智算系统的整体稳定性；在安全方面，运用曙光提供的全套解决方案，实现从底层硬件到上层运维服务的全栈安全可靠；计算服务平台赋能智算中心实现从软硬件部署到应用价值落地，目前平台上线了模型仓库、开发工具及智能应用，帮助智算中心提升运营能力。同时，公司积极参与智算中心行业标准建设，基于在智能计算领域的实践成果，为中国信息通信研究院《智能计算中心总体技术要求》提供关键的技术支持和专业建议。

与此同时，曙光5A级智算中心遵循开放包容的建设原则，一方面积极完成各类主流算法、框架及大模型的适配优化工作，显著降低算力应用门槛，使得AI技术触手可及；另一方面，鼓励上下游合作伙伴深度参与项目建设，共同构建繁荣共生的产业链条，激发创新活力。

目前，曙光5A级智算中心已相继在合肥、长沙、北京、青岛等地投入运营，其强大的算力引擎正源源不断赋能金融、通信、能源、工业、医疗、科研等多元领域，催生出海量生态应用落地，有力驱动各地区、各行业步入数智化发展的快车道。中科曙光以实际行动诠释了智能计算的革新力量，持续为我国数字经济高质量发展注入新动能。

在建设和运营先进算力中心的基础上，曙光进一步布局建设了“全国一体化算力服务平台”，通过网络链接各类算力基础设施资源，并建立以应用服务为导向的创新型算力服务平台，降低用户使用门槛、提高算力应用水平，为千行百业提供“集算力、数据、应用、运营、运维为一体的服务”。2023年，子公司曙光智算深度参与行业标准和科技创新，通过中国信息通信研究院“可信算力服务-智算平台”评测，成为首批通过的4家厂商之一；2023年2月，曙光智算联合宁夏回族自治区打造首个一体化算力交易调度平台；6月，公司作为“北京算力互联互通验证平台”参建与接入服务商参与二期成果发布，并被授予首通调度试验成功企业与主要贡献专家、算网云开源操作系统项目首批示范单位等荣誉；10月，曙光智算AC平台获评深圳国际人工智能展“最佳人工智能服务平台”奖项。

深化产业链上下游合作

在数字经济、人工智能等领域先进技术高速发展的新时代，上下游、软硬件厂商合作更频繁、更紧密。曙光公司积极在产业链上下游开展生态建设工作，依托在计算领域的先发优势和技术底蕴，紧密围绕中国本土市场需求，携手产业链上下游众多合作伙伴，通过协同研发、融合创新、资本投资，大力开展技术研究、标准制定、系统适配、测试评估、生态建设、人才培养等方面的合作。

2024年，曙光公司加快构建开放协同的生态体系，推进构建软硬件相互促进、相互带动的协同发展模式，加强公司核心产品与产业链厂商的适配；同时，加强产学研协同创新，推进各方在技术、应用、市场等维度的优势互补和资源共享，在研发领域，继续加大与人工智能、存储、网安等行业厂商的技术合作；与合作伙伴共享优势行业，加大公司核心产品在其他行业的拓展能力；在基础设施运营建设方面，与生态伙伴基于产品互补打造联合解决方案，共建数字经济的标杆项目。在行业发展的关键时期，公司将与更多产业伙伴合力前行，一同构建起千行百业数智化转型中的关键力量，帮助更多用户在数字时代塑造关键竞争力。

5.2 广州数字科技集团

广州数字科技集团有限公司以电子信息制造业起家，是我国最早布局的军工电子骨干企业，深耕电子信息制造业近70多年，连续38年入选中国电子信息百强。近年来，集团充分利用国有资本投资公司平台和电子信息制造业的基础优势，加快推动数字经济与实体经济深度融合，逐步形成以人工智能和数字经济为核心的多元化产业格局，技术、产品、服务和解决方案涉及数字经济、数字社会、数字政府和数字生态各个方面。

广电运通创立于1999年，是广州数字科技集团培育出来的第一家A股上市公司，也是集团行业人工智能业务板块的主力军。广电运通从国内金融自助设备起步，连续多年位居国内第一、国际前三。公司拥有由4名院士领衔，包括博士、硕士在内的超2600人的专业研发团队，布局算法、算力、数据、场景四大人工智能要素协同发展，以aiCore System数字技术底座为基石，深度挖掘数据要素价值，打造“望道”行业大模型，有效打造了数字经济优势产业集群。

2024年2月，广州无线电集团改组为广州数字科技集团，肩负起推动广州数字经济高质量发展的重任。另一方面，广州数字科技集团成立二级企业广州数据集团，负责广州人工智能公共算力中心建设运营与公共数据开发利用等工作。广电运通依托控股股东改组契机，充分发挥与广州数据集团在智算、数据要素领域业务上的协同作用，通过控股广电五舟、承建算力中心、构建数据要素产业链等重大举措，全面构筑公司核心竞争力，打造新的业绩增长点。值得一提的是，公司持续加码智算、数据要素的投入力度，计划将现所在的科学城产业园改造为数据产业园，推进数据服务商引入，提供适应数字经济业务和数据业务发展的新型基础设施，构建多元发展的数据要素生态。

控股广电五舟，深化国产算力布局

自人工智能战略升级以来，广电运通加速在通用算力和人工智能算力基础设施板块的能力建设，于2021年入股广电五舟，成为其第一大股东。经过四年多的协同发展，广电运通于2024年5月召开董事会审议通过了《关于收购广州广电五舟科技股份有限公司股权的公告》，拟进一步收购广电五舟10%的股份，并通过重要股东表决权委托协议方式成为广电五舟的控股股东。

广电五舟是一家专业的智能算力产品和解决方案提供商，以服务器的国产化、差异化和软硬一体化为核心业务，是国产服务器行业定制领域的领导品牌，同时也是华为的鲲鹏整机合作伙伴和昇腾AI战略伙伴，主要产品包括服务器整机、金融工控机、模块化网安平台、云终端、分布式存储一体机、超融合软件、高性能计算集群管理系统等一系列全面的软硬件算力产品。2023年，受益于国产算力产业的深入推进，广电五舟实现营业收入12.98亿元，同比增长133%，整体发展态势良好。

算力是数字经济时代的新质生产力，随着国产服务器市场渗透率逐步提高，广电运通将紧抓国产算力市场机遇，不断构建和完善数字技术及产品体系，强化市场竞争力，在信创产业链中发挥更大作用，为高质量发展注入新活力。



资料来源：广电运通

共建“鲲鹏+昇腾”项目，构筑广东计算产业生态

2020年3月，在市委市政府统一部署下，广州数字科技集团与华为签署合作协议，宣布成立“鲲鹏+昇腾”生态创新中心，联合打造广州“鲲鹏+昇腾”高性能计算产业生态。

基于华为“鲲鹏”和“昇腾”处理器，广电运通开始进行AIoT的战略算力布局。比如在服务器上叠加广电运通的AI大数据平台aiCore System, 加速拓展和复制人工智能技术应用场景，在其已拥有广泛客户基础的政府、银行、公安、交通、财税及文化旅游等领域打造全栈式自主可控技术体系。目前，广电运通在“鲲鹏+昇腾”的业务，已实现了金融、党政、安防、教育、医疗、交通等多行业的场景落地，为打造全国领先的计算产业集聚区、适配样板区、应用示范区提供了有力支撑。

“鲲鹏+昇腾”计算产业是中国信创产业生态的关键一环，对于推动国产软硬件的发展和增强信息安全具有重要意义。未来，广州数字科技集团、广电运通将秉承国家战略，与生态伙伴共建“鲲鹏+昇腾”计算产业生态，助力广州、广东乃至中国新一代信息技术产业的加速发展，为数字基建，为产业链供应链的现代化贡献更大力量。

卡位大湾区智算中心建设，部署算力“新基建”

粤港澳大湾区作为我国开放程度最高、经济活力最强的区域之一，拥有得天独厚的区位优势、产业基础和创新能力。数字经济的发展，催生智能算力建设提速。按照市委、市政府的统一部署，广州数据集团旗下广州广电研究院有限公司建设运营广州人工智能公共算力中心（简称“算力中心”），通过“一中心五平台”，对外提供人工智能普惠算力服务、应用创新孵化服务、产业融合赋能服务、科技协同创新服务和人才培养服务，构建数字经济全要素发展体系，助力广州数字经济发展。

算力中心承建采用公开比选方式。2022年2月，广电运通中选算力中心一期项目，金额3.67亿元，目前已正式投入使用。2023年底，公司中选算力中心二期项目，金额4.73亿元，目前正在承建中。算力中心采用公司自主研发的算力运营平台，成为全国“企业投资、政府支持”的人工智能产业基础共性技术平台，构建“算力服务、科技创新、应用培育、产业赋能”的AI全链条闭环，实现技术、算力、数据、市场多方融合，构建“产学研用”全过程创新链，已建成的算力中心一期项目资源使用率达70%—80%。

除了专注算力能力建设，广电运通也注重云能力的培育。目前，公司围绕“四云两中心”开展自主研发，打造广州政务信创云、公安信创云、国资国企云、AI智算云等专属云服务。2024年5月，广电运通全资子公司运通信息中标广州市数字政府“一朵云”项目，中标金额5.75亿元，实现了在云服务领域的重大突破。



图60广电运通积极参建智算中心

资料来源：广电运通

另外，广电运通通过参股广州数据交易所、建设广州人工智能公共算力中心，收购大数据征信企业中数智汇等举措，全面布局数据要素业务，构建了数据价值化的全链条，在数据加工、运营、交易等环节持续打造核心竞争力。随着公司在数据要素全链条业务服务能力上的逐步完善，已在广东率先取得了数据要素业务的阶段性进展。

在数字化浪潮席卷全球，数字中国建设如火如荼的时代背景下，广州数字科技集团、广电运通将推进算力、数据要素走实走深，用数字力量开启更多领域的“数字大门”，打造新质生产力，奔向数字经济新时代。

六、智算发展展望与政策建议

6.1 加强智算资源整合与布局优化

随着人工智能技术的迅速发展，对智能算力需求的大幅增加，智算中心的建设需要加强多元异构智算资源整合和整体布局优化，以提供稳定高效的算力供给。首先，加快推动异构、异属、异域智能算力资源整合并网，通过政策保障和资金支持激励各地智算中心加入全国一体化算力网，加强多元异构算力互联互通和统一服务，探索跨区域算力调度体系，实现供需之间的灵活对接和资源的统一管理，提升智算资源整体利用率。其次，加强智算中心整体布局优化，结合“东数西算”工程，加强东西部地区智算产业协同建设，充分发挥西部地区丰沛的绿电资源优势，将模型训练等对算力需求量大、时延敏感度低的人工智能应用场景向西逐步迁移，补强西部地区产业发展短板，释放东部地区能耗压力，谋划全国智算产业一盘棋。

6.2 加快培育国产软硬件生态

推动国产软硬件厂商开放协同，实现芯片、算法、模型、应用的融合创新。在国产硬件方面，面向人工智能云端分布式训练需求开展高性能通用芯片研发，面向边缘端应用需求开展低时延、高能效专用芯片研发，积极引导智算中心应用国产人工智能芯片，加快提升人工智能算力供给的国产化率，推动自主可控软硬件算力生态建设，促进国产芯片技术的快速迭代和升级。在国产软件方面，重点支持在高端人工智能芯片、异构计算、分布式训练、人工智能应用等方面技术实力雄厚的企业构建开源生态，开发高效易用的人工智能芯片编译器、加速库、工具链、深度学习框架等基础软件，提供高质量的开源服务，引领并吸引各方积极参与贡献力量。

6.3 促进产业链协同与创新升级

智算中心建设除满足大模型训练等特定场景需求外，还应面向产业智能化发展需求，辅助构建开放融合的人工智能生态，助推AI产业化与产业AI化转型发展。一方面，打造开放合作的智算服务社区和产业联盟，构建智能算力公共服务平台，充分利用市场化机制在资源高效配置方面的能力，降低技术壁垒和市场准入门槛，推动产业链上下游企业的协同发展。另一方面，培育政产学研金用联合创新体系，推动高校、科研院所和企业的创新联合攻关，打造一批典型应用示范场景，推动科研成果转化落地，促进产业链创新链融合发展。

6.4 加强智算中心运营保障

在运营机制方面，构建政府引导、市场驱动的智能算力运营服务体系，发挥政府统筹引导作用，积极对接各类算力供需主体，坚持市场驱动，激发各类市场主体积极性，促进算力供需精准对接，实现智算产业的可持续发展。在人才培养方面，开展校企联合实训，加强智算中心和人工智能相关设计、运维、管理和创新应用的人才队伍建设，注重实践应用与产业需求相结合，提升人才在具体业务场景中的解决实际问题能力。在资金保障方面，鼓励金融机构创新金融服务和融资产品，加大金融机构信贷支持力度，建立以政府投资为引导，以企业投资为主体，金融机构积极支持，民间资本广泛参与的智算产业发展投融资模式。