

2023 智能算力发展白皮书

新华三集团
中国信息通信研究院
2023 年 8 月

版权声明

本白皮书版权属于新华三集团和中国信息通信研究院,并受法律保护。转载、摘编或利用其它方式使用本白皮书文字或者观点的,应注明“来源:《2023 智能算力发展白皮书》”。违反上述声明者,将追究其相关法律责任。

前言

发展数字经济是把握新一轮科技革命和产业变革新机遇的战略选择，在数字化、智能化时代，算力就是数字经济发展的核心生产力，智能算力就是创新力。当前人工智能正向多场景、规模化、融合度高的阶段发展，数据量急剧增长，算法模型愈加复杂，应用不断延伸，这对智能算力的发展提出了更高要求。放眼世界，很多国家都在积极开发和部署智能算力资源，以塑造未来发展优势。在这样的背景形势下，新华三集团联合中国信息通信研究院，共同编制了《2023 智能算力发展白皮书》，希望与业界同仁共同推进我国智能算力高质量发展。

白皮书提出了智能算力内涵定义，对当前全球及我国智能算力的总体情况、智能算力应用及技术发展现状进行系统性梳理。同时，深入分析了智能算力发展面临的挑战，并提出解决方案。最后白皮书展望智能算力未来发展趋势，并提出下一步发展建议。

时间仓促，白皮书不足之处在所难免，我们后续将不断更新完善，意见建议请联系 dceco@caict.ac.cn。

目 录

1 背景与意义	1
1.1 智能算力的定义与内涵	1
1.2 发展智能算力的意义	2
2 智能算力总体情况	3
2.1 全球智能算力总体情况	3
2.2 我国智能算力总体情况	4
2.3 我国智能算力行业应用分布	5
3 智能算力发展现状和挑战	6
3.1 智能算力技术层面应用发展现状	6
3.2 智能算力应用层面发展现状	12
3.3 智能算力发展挑战及解决方案	22
3.3.1 算力需求	23
3.3.2 能耗	23
3.3.3 算法复杂度	24
3.3.4 数据隐私和安全	25
3.3.5 生态合作	26
4 智能算力未来发展趋势	27
4.1 人工智能加速渗透，多样化场景催生多元化算力需求	27
4.2 政策驱动，智能算力低碳发展成硬性要求	28
4.3 边缘智能应运而生，边缘计算与人工智能融合发展	28
4.4 智算中心建设加速，应对高质量算力需求	29
4.5 模型规模不断扩展，海量多元化数据亟需巨量化算力	30
4.6 自主学习能力提升，推动算力实现更高层次智能	30
5 智能算力发展展望与建议	31
5.1 智能算力发展展望	31
5.2 智能算力高质量发展建议	33

图 目 录

图 1	近两年全球算力规模情况	4
图 2	近两年我国算力规模情况	5
图 3	中国人工智能行业渗透度	6
图 4	深度学习发展历程	9
图 5	计算机视觉产业链全景图谱	11
图 6	元宇宙虚拟世界表示、交互方式及内容创作方式演变	14
图 7	国内外 AIGC 产业化情况	15
图 8	数字孪生的技术体系	17
图 9	边缘智能布局架构	18
图 10	我国智能算力规模及预测	32

1 背景与意义

1.1 智能算力的定义与内涵

全球算力发展正面临应用多元化、供需不平衡等挑战,人工智能、数字孪生、元宇宙等新兴领域的崛起,推动算力规模快速增长、计算技术多元创新、产业格局重构重塑,智能算力作为数字经济时代新的生产力,对推动科技进步、赋能行业数字化转型以及经济社会发展发挥着日益重要的作用。智能算力即人工智能算力,是面向人工智能应用,提供人工智能算法模型训练与模型运行服务的计算机系统能力。智能算力通常由 GPU (Graphics Processing Unit, 图形处理器)、ASIC (Application Specific Integrated Circuit, 专用集成电路)、FPGA (Field Programmable Gate Array, 现场可编程逻辑门阵列)、NPU (Neural network Processing Unit, 神经网络处理器) 等各类专用芯片承担计算工作,在人工智能场景应用时具有性能更优、能耗更低等优点。

智能算力是数字经济时代的重要支撑。数字经济依赖于数据的处理和分析,而智能算力为这些操作提供了强大的支撑。企业和个人利用智能算力提供的高性能计算能力处理海量的数据,实现快速、准确的数据分析,从而为企业的决策和发展提供更多的信息和支持。人工智能、大数据、物联网等新兴技术在智能算力支持下能够更加高效地进行数据处理、模型训练和决策推断,加速技术落地,推动数字经济与实体经济深度融合。

智能算力是人工智能发展的动力。智能算力使得运算速度大幅增

加，处理复杂数据的能力大幅提升，传统的人工任务逐渐被自动化和智能化取代，人们开始寻求更加复杂和高级的任务，如自动驾驶、自动翻译、半自动化医疗，人工智能领域进入了一个全新的阶段。另外，随着智能算力的提升，新的算法和技术不断涌现，为人工智能发展带来了新的创新机遇。通过大规模并行计算、深度学习、神经网络等技术，在智能算力的支持下，人工智能在各个领域都得到了广泛应用，推动了人工智能技术的进一步突破。

智能算力是科技创新的新引擎。智能算力为科技创新提供强大的计算支持，促进了科技创新的效率和质量。学术和工业界能够处理更加复杂的计算任务，科研人员可以更快速地进行大规模数据处理、模拟实验和模型训练，极大地提高了科技创新的效率和质量。另外，智能算力可以推动新兴技术的突破，催生众多的创新应用，为科技的发展带来新的思路和方法。在人工智能技术当中，智能算力是算法和数据的基础设施，更快速、更高效的数据处理能力使得人工智能可以应用于更多的领域。

1.2 发展智能算力的意义

智能算力作为关键生产力要素，推动数字经济高速发展。智能算力使得数据的处理和分析变得更加高效、准确，使得庞大的数据量可以更加高效地被挖掘和利用，为数字经济提供了强大的基础支持。如在制造业领域，智能算力可以提供实时的生产数据，实现智能化的生产管理，提高企业的生产效率和产品质量。在服务业领域，智能算力可以通过大数据的分析和挖掘，实现自动化的客户服务，提供更加智

能、个性化的服务体验。此外，智能算力也推动了数字经济与传统产业的融合，通过与人工智能、云计算、物联网等技术的结合，加速我国实体经济加速向数字化、网络化、智能化方向转变。

智能算力为人工智能发展提速，促进行业应用。目前，人工智能技术高速发展，智能化场景在行业的落地随着时间的推移，正呈现出更加深入、更加广泛的趋势，对智能算力的需求与日俱增。人工智能与制造、交通、医疗、农业等各领域融合日益深入，持续推动质量变革、效率变革、动力变革，源源不断地为经济高质量发展提供新动能。未来五年，随着人机交互、机器学习、计算机视觉、语音识别技术的成熟，人工智能将在企业市场中加快应用与落地，赋能传统行业转型升级，而智能算力将助力于人工智能的持续快速发展。

智能算力为科技进步提供新动力，推动科技跨越式发展。智能算力为国家创新力的发展带来实质性推进，不仅在应用科学的突破上发挥了重要作用，也开始渗透到基础科学领域。科学家们越来越多地利用人工智能技术，从数据中建立模型，重点围绕新药创制、基因研究、新材料研发等领域加速对前沿科学问题的探究。如科研领域利用人工智能进行蛋白质折叠体结构的研究、抗菌耐药性基因的检测和识别；医药领域 AI 计算辅助疫苗和药物研发用于靶点选择和验证先导化合物筛选和优化等研发环节，从传统“手工试错”向计算辅助模式转变最大化缩短研发周期。

2 智能算力总体情况

2.1 全球智能算力总体情况

全球智能算力的总体情况呈现快速增长的趋势。截至到 2022 年底，全球算力总规模达到 650 EFLOPS，其中，通用算力规模为 498 EFLOPS，智能算力规模为 142 EFLOPS，超算算力规模为 10 EFLOPS。智能算力规模与去年相比增加了 25.7%，规模占比达 21.9%。IDC 预测，全球 AI 计算市场规模将从 2022 年的 195.0 亿美元增长到 2026 年的 346.6 亿美元。

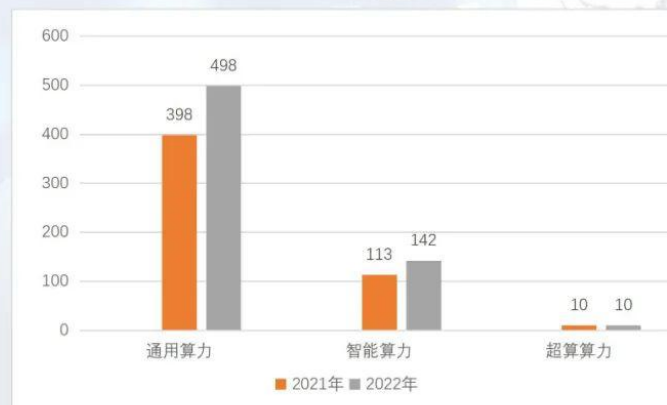


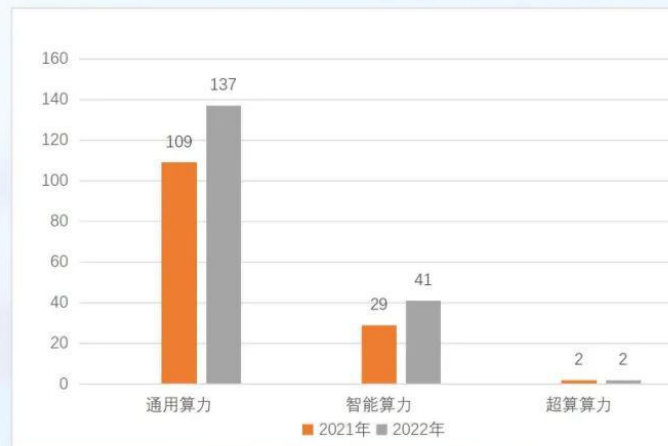
图 1 近两年全球算力规模情况¹

(数据来源：Gartner、IDC，中国信息通信研究院整理)

2.2 我国智能算力总体情况

在算力规模方面，截止到 2022 年底，我国算力总规模为 180 EFLOPS，排名全球第二。其中，通用算力规模为 137 EFLOPS，智能算力规模为 41 EFLOPS，超算算力规模为 2 EFLOPS。中国智能算力正处于高速增长阶段，智能算力规模与去年相比增加了 41.4%，规模占比达 22.8%，，超过全球整体智能算力增速（25.7%）。

¹ 部分数据来源：《中国算力白皮书（2022 年）》

图 2 近两年我国算力规模情况²

(数据来源:《中国算力白皮书》、中国信息通信研究院整理)

2.3 我国智能算力行业应用分布

人工智能在各行业应用程度均呈现不断加深的趋势,应用场景越来越广泛。智能算力在行业应用情况可根据人工智能的行业渗透度来分析,与 2021 年相比,各行业人工智能渗透度明显提升。其中,互联网行业依然是人工智能应用渗透度和投资最高的行业;金融行业的人工智能渗透度从 2021 年的 55 提升到 62,智能客服、实体机器人、智慧网点、云上网点等成为人工智能在金融行业的应用典型;电信行业的人工智能渗透度从 2021 年的 45 增长到 51,人工智能技术融入电信网络的构建、优化,并为下一代智慧网络建设提供支撑;制造业的人工智能渗透度从 40 增长到 45,结合人工智能技术的传统制造业的智能化改造,已成为产业升级的热点。

² 部分数据来源:《中国算力白皮书(2022 年)》

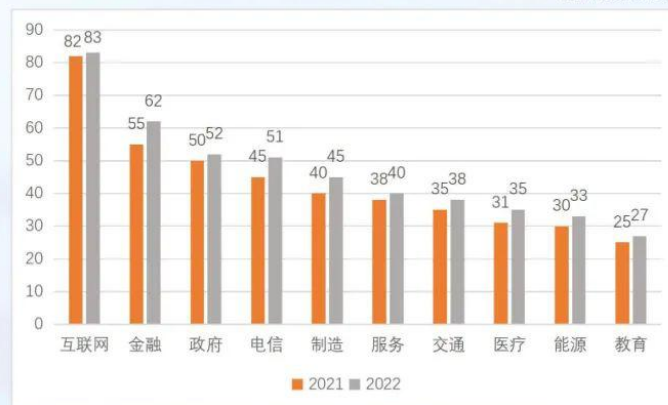


图3 中国人工智能行业渗透度

(数据来源: IDC, 2022)

3 智能算力发展现状和挑战

3.1 智能算力技术层面应用发展现状

人工智能技术高速发展,应用方向逐渐多样化和复杂化。智能算力主要有三个优点,一是能够提供大规模数据处理和复杂计算的能力,满足人工智能算法对于高性能计算的需求;二是能够加速人工智能模型的训练和推理过程,提高算法的效率和准确度;三是能够与其他技术手段结合,如云计算、大数据分析和边缘计算,实现人工智能在各行各业的广泛应用。智能算力满足人工智能高并发、高弹性、高精度的计算需求,推动人工智能技术的不断升级与应用。高性能的计算能力为机器学习、深度学习和自然语言处理等人工智能技术的发展提供了有力的支持,通过智能算力的支持,人工智能算法能够处理和分析大规模的数据,实现复杂任务的智能化。未来五年,人工智能将在企业市场中加快应用与落地,智能算力将成为创新的核心推动力。

3.1.1 机器学习

机器学习 (Machine Learning) 本质是通过计算机从大量的数据中找到整合数据的规律, 从而实现对于数据未来走向的预测。由机器学习算法支撑的机器视觉、听觉和语音交互被应用各种产品和服务中, 进而带来了 AI 在商业应用方面的爆炸式增长。目前通过让机器从大量数据中自主学习, 机器学习使计算机具有了更强大的智能和能力, 已经被广泛应用于图像识别、语音识别、医疗诊断、金融风控、智能推荐等领域。同时, 机器学习也开始参与到了计算机内部体系的研究和设计过程中, 例如在计算机的翻译器、硬件处理器以及软件工程等设计开发方面利用更加现代化的编程语言。

智能算力在机器学习中的作用主要是为深度学习模型训练、大规模数据处理、实时数据分析与预测、自动化模型选择和调参、分布式机器学习等提供强大的计算力。在机器学习中, 通常需要处理大规模的数据集。例如, 图像分类任务中, 需要处理成千上万张图像来训练和测试模型, 智能算力提供了并行计算和分布式计算的能力, 可以快速处理大规模数据, 加速训练过程。智能算力还可以在实时数据分析和预测方面发挥作用。例如, 通过将机器学习模型部署在智能算力的环境中, 可以实时地监测和分析海量的数据, 并利用模型进行实时预测和决策。

3.1.2 深度学习

深度学习 (Deep Learning) 是一种基于多层神经网络的机器学习方法, 主要特点是能够处理复杂的非线性问题, 可以学习和理解图像、

声音和自然语言等复杂数据，并具有优秀的预测和决策能力。2011年，微软和谷歌率先将深度学习技术应用于语音识别，大大提升了识别率；2012年，深度学习开始用于图像识别，在 ImageNet 数据集上将原有识别错误率降低了百分之十一；同年，微软公开了采用深度学习技术的“全自动同声传译系统”，该系统几乎能达到和人一样的水平，实现实时翻译；2013年，百度宣布成立深度学习研究所，专注于该技术的研究；2016年，谷歌开发的人工智能 AlphaGo 战胜专业围棋选手，这一成果迎来了深度学习的热潮。目前，应用较广泛的深度学习框架有 TensorFlow、Caffe、Theano、MXNet、Torch、PyTorch 等，实际应用主要有计算机视觉、语音识别、语言处理等。

随着深度学习的推进，人工智能逐渐应用到各个领域，对算力的需求越来越高，其核心是人工智能要达到目标必须不断地进行大规模、高频次的数据训练，经过训练神经网络才能总结出规律，对新的样本才能进行判断和分析。智能算力在深度学习中的应用现状主要体现在以下几个方面：一是训练模型，深度学习模型具有复杂的结构和大量的参数，需要大量的计算资源进行训练，智能算力通过 GPU、TPU 等高性能的计算设备，加速深度学习模型的训练过程。二是推断推理，智能算力通过高性能计算设备和专门的推理芯片加速深度学习模型的推断过程，提高了模型的实时性和稳定性。三是模型优化，通过智能算力可以对模型进行自动化的超参数调优、网络结构搜索、模型剪枝等操作，进一步提高模型的精度和效率。

推理 时期	1960s	赋予机器逻辑推理能力使机器获得智能:当时的AI程序证明力一些著名的数学定理,但由于缺乏知识远不能实现真正的智能
知识 时期	1970s	将人类的知识总结出来教给机器使机器获得智能即“专家系统”,在很多领域获得大量进展,但由于人类知识量巨大,故出现“知识工程瓶颈
机器学习 时期	1980s	连接主义较为流行,代表方法为神经网络
	1990s	统计学习占据舞台,代表方法包括支持向量机等
	21Cent.	深度神经网络被提出,连接主义卷土重来。随着数据量和计算能力的不断提升,以深度学习为基础的诸多AI应用逐渐成熟

图 4 深度学习发展历程

3.1.3 自然语言处理

自然语言处理（Natural Language Processing，NLP）指的是基于人类日常交流过程当中所使用的自然语言与计算机进行交互的一种技术类型，涵盖语言学、计算机科学、数学、新闻学等一系列学科内容，是计算机领域以及人工智能领域未来发展的重要方向。针对特定应用，具有相当自然语言处理能力的实用系统已经出现，甚至开始产业化。如多语种数据库和专家系统的自然语言接口、各种机器翻译系统、全文信息检索系统、自动文摘系统等。在人工智能技术的支持下，自然语言处理系统的适应能力不断提升。

智能算力在自然语言处理中的应用主要体现在以下几个方面：一是语言模型，基于深度学习的语言模型（如 BERT、GPT 等）极大地提升了 NLP 任务的性能，这些模型能够学习到丰富的语义信息和潜在语言规律，使得计算机能够更好地理解和生成自然语言文本。二是

机器翻译，神经机器翻译（NMT）技术已经成为主流，NMT 模型能够将一种语言翻译成另一种语言，实现更准确和流畅的翻译质量。三是问答系统，通过结合 NLP、信息检索、文本匹配和语义分析等技术，问答系统能够根据用户提出的问题返回准确的答案。四是文本分类与情感分析，NLP 技术可以利用深度学习方法对文本进行分类和情感分析，广泛应用于舆情监测、电商评论分析等领域。

3.1.4 计算机视觉

计算机视觉是指让计算机和系统能够从图像、视频和其他视觉输入中获取有意义的信息，并根据该信息采取行动或提供建议。计算机视觉技术主要应用包括图像分类和目标检测、人脸识别、图像生成和风格迁移、三维重建和虚拟现实、视频分析和行为识别等。计算机视觉硬件方面主要还是由国外巨头把控，国内市场份额不足，特别是芯片领域和算法算力方面，但在数据方面，我国市场巨大，应用广泛；中游技术层面部分技术我国已处于领先地位，如人脸识别算法方面、物体检测技术；下游应用层面成果广泛，我国已形成了全面布局行业解决方案，特别在智慧安防、智慧金融、互联网领域市场增长迅速，颇具竞争优势。智能算力在计算机视觉中的主要应用是使用深度学习模型，特别是卷积神经网络，进行图像分类、目标检测和图像分割等任务，这些模型在智能算力支撑下经过大规模数据的训练，能够准确地识别和解析图像中的内容。



图5 计算机视觉产业链全景图谱

3.1.5 数据分析和挖掘

数据分析和挖掘技术是从大规模数据中提取有价值信息的重要工具，主要通过统计、计算、抽样等相关的方法，来获取基于数据库的数据表象的知识。随着大数据时代的到来和技术的不断进步，数据分析技术和工具不断涌现，包括数据可视化和探索、机器学习和深度学习、数据挖掘算法、异常检测和异常数据分析、大规模数据处理和分布式计算等。这些工具和技术的出现及应用，提高了分析数据的效率和精度，并且增加了数据科学家对数据解释的可信度。目前，解决问题和做决策所需的数据已经不再是少数行业，广泛应用于金融、教育、环境和安防等各个领域，未来数据分析技术及其工具将被广泛应用并将向自动化、智能化发展。

智能算力在数据分析和挖掘领域的主要应用是使用机器学习算法，如决策树、支持向量机、随机森林等，对大规模数据进行模式和关联性的挖掘。另外也可以应用于深度学习模型，如神经网络，对大

量的结构化和非结构化数据进行高级分析和挖掘，如图像、文本和语音数据等。另外智能算力可以帮助实现自动化的数据处理和特征工程。通过智能算力的高效计算和高度并行的能力，可以加速大规模数据的清洗、转换和特征提取过程。

3.2 智能算力应用层面发展现状

智能算力推动人工智能技术落地，算力释放成为生产力。人工智能技术的核心是模型训练与推理，而对于庞大的数据集和复杂的算法模型，需要大量的计算资源来支持。智能算力可以更快、更高效地进行模型训练和优化，从而加速人工智能技术的发展，推动人工智能技术应用于更多的领域和场景，为社会和企业创造实实在在的价值。例如，在制造业，智能算力可以优化供应链管理与生产流程，实现智能制造；在金融领域，智能算力可以提供更准确的风险评估与投资建议，提高金融机构的决策能力；在汽车行业，将计算机视觉和机器学习与GPS定位技术、传感器技术、大数据技术等进行有机融合，为汽车的自感知、自学习、自适应和自控制提供支持。如今算力被视为生产力，成为传统产业转型升级的重要支点，积极释放数据要素的创新活力，赋能各行各业。

3.2.1.基础应用

1.元宇宙

元宇宙（Metaverse）是人类运用数字技术构建的，由现实世界映射或超越现实世界，可与现实世界交互的虚拟世界，具备新型社会体系的数字生活空间，集成了一大批现有技术，包括5G、云计算、人

工智能、虚拟现实、区块链、数字货币、物联网、人机交互等。元宇宙中的虚拟世界构建经历了从文本到 2D 再到 3D 的形式演变，交互方式也由命令行转变为图形界面再到最近的虚拟现实、增强现实以及混合现实等更加自然的方式。在内容创作上，从早期的专业创作内容逐步过渡到用户创作内容的形式，且有望在未来进入人工智能创作内容的范式。2021 年新冠疫情的推动下元宇宙爆火，Soul App 首次提出构建“社交元宇宙”，微软打造“企业元宇宙”；同年，英伟达推出全球首个为元宇宙建立提供基础的模拟和协作平台，美国脸书（Facebook）宣布更名为“元”（Meta），来源于“元宇宙”（Metaverse），并宣布两年内对 XR 投入 5000 万美元。

智能算力为元宇宙的构建提供了强大的计算支持。通过云计算、分布式计算、边缘计算等技术，智能算力能够快速有效地处理海量数据和复杂计算任务，实现元宇宙中的虚拟现实、人工智能、物联网等应用。在虚拟现实方面，智能算力可以支持逼真的图形渲染、物理模拟和实时交互，提供沉浸式的虚拟体验；在人工智能方面，智能算力能够训练和优化复杂的神经网络，实现自动化的语音识别、图像处理和情感分析。未来智能算力将与边缘计算、区块链等新兴技术相结合，更好地满足元宇宙应用对于大规模数据处理、实时互动和高度智能化的需求。



图6 元宇宙虚拟世界表示、交互方式及内容创作方式演变

2.AIGC

AIGC (AI-generated content, 人工智能生成内容技术) 是基于 GAN、预训练大模型、多模态技术融合的产物, 通过已有的数据寻找规律, 并通过泛化能力形成相关内容。随着 ChatGPT 的爆火, 整个 AIGC 领域瞬间被点燃, 绘画、建模、视频、影视等领域纷纷探讨应用的可能性, 在 OpenAI 的 ChatGPT 之外, 谷歌、百度、阿里、字节等大厂纷纷投入各类大模型的研发中, 属于 AIGC 的大时代已经来。AIGC 已经能够生成文字、代码、图像、语音、视频、3D 物体等各种类型的内容和数据, 多模态技术的成熟让 AIGC 可应用的广度不断扩展, 未来应用潜力更大。

从技术角度来看, AIGC 的背后是算力、数据、算法等核心要素的有机融合, 模型越大, 对算力要求越高。ChatGPT 爆火的背后, 本质上是人类在 AI 领域软件(数据、算法)、硬件(算力)综合能力大幅提

升以后的一次爆发式体现。基于飞天智算的阿里云深度学习平台 PAI，将计算资源利用率提高 3 倍以上，AI 训练效率提升 11 倍，推理效率提升 6 倍；新华三推出专门为大模型训练而生的 AI 服务器及 51.2T、800G CPO 硅光数据中心交换机，支持大算力调度的傲飞算力平台；“文心一言”背后的算力基础设施是百度智算中心，是亚洲最大的单体智算中心，可承载约 28 万台服务器，算力规模达 4EFLOPS。未来 AIGC 对智能算力的需求将更加强劲，GPU、FPGA、ASIC 等底层硬件中包含的智能算力价值将被重塑。

图 7 国内外 AIGC 产业化情况³

3. 数字孪生

数字孪生（Digital Twin，数字双胞胎），是充分利用物理模型、传感器更新、运行历史等数据，集成多学科、多物理量、多尺度、多概率的仿真过程，在虚拟空间中完成映射，从而反映相对应的实体装备的全生命周期过程。国外关于数字孪生的理论技术体系较为成熟，当前已在相当多的工业领域实际运用。国内数字孪生技术处于起步阶

³ 来源：中国信息通信研究院整理

段,研究重点还停留在理论层面。数字孪生技术目前呈现出与物联网、3R(增强现实(Augmented Reality,AR)、虚拟现实(Virtual Reality,VR)和混合现实(Mixed Reality,MR))、边缘计算、云计算、5G、大数据、区块链及人工智能等新技术深度融合、共同发展的趋势。

智能算力可以支持数字孪生模型的建模、仿真和优化并推动其在行业中的广泛应用。从技术角度来看,通过云计算、大数据分析和机器学习等技术手段,智能算力能够处理和分析大规模的数据,并生成高度精确的数字孪生模型。同时,智能算力还能够实现实时的数据同步和模型更新,提高数字孪生系统的性能和可靠性。从行业应用角度来看,智能算力在数字孪生领域已经得到广泛应用。在制造业领域,智能算力能够建立物理系统的数字孪生模型,并通过数据监测和算法优化提高生产效率和质量;在城市规划和交通管理方面,智能算力能够建立城市的数字孪生模型,优化交通流量和环境布局;在医疗领域,智能算力能够创建人体的数字孪生模型,辅助手术规划和医学研究。



图8 数字孪生的技术体系

4.边缘智能

边缘计算和人工智能彼此之间相互赋能并催生了新的研究领域——边缘智能。边缘计算将计算、网络、存储等能力扩展到物联网设备附近的网络边缘侧，而以深度学习为代表的人工智能技术让每个边缘计算的节点都具有计算和决策的能力，这使得某些复杂的智能应用可以在本地边缘端进行处理，满足了敏捷连接、实时业务、数据优化、应用智能、安全与隐私保护等方面的需求。在边缘智能中，边缘计算为人工智能提供了一个高质量的计算架构，对一些时延敏感、计算复杂的人工智能应用提供了切实可行的运行方案。目前，边缘智能正深入推动智慧交通、智能制造、云游戏等应用的发展，促进了产业的实现与落地，为全面提升智能化水平提供了重要保障。



图 9 边缘智能布局架构⁴

5.推荐系统

推荐系统是一种信息过滤系统，可以很好地解决信息过载问题，提高信息利用率，是整个推荐过程的关键部分，根据推荐需求，结合建立的用户模型，将最符合用户偏好的物品推荐给用户。推荐系统应用于各行各业，推荐的对象包括：电影、音乐、新闻、书籍、学术论文、搜索查询、分类、以及其他产品。也有一些推荐系统专门为寻找专家、合作者、餐厅、美食、金融服务、生命保险、网络交友，以及Twitter 页面设计。人工智能在推荐系统中的应用已经非常广泛，通过分析 and 理解用户的行为和兴趣，可以为用户提供个性化的推荐服务，提高用户满意度和平台的业务价值。主要应用方面包括协同过滤、深度学习、强化学习、自然语言处理等方面。

智能化是推荐领域不可避免的趋势，目前推荐系统与深度学习、

⁴ 来源：《边缘计算技术白皮书》，ODCC

知识图谱的结合也成为了业界研究的重点。例如基于深度学习的推荐，深度学习的优势在于表达能力强，能挖掘出更多数据中潜藏的模式，并且结构十分灵活，可根据不同推荐场景或不同特点的数据来进行调整。通过人工智能技术以及大数据的支撑，对购买意向、阅读习惯、浏览需求等信息进行学习和预测性分析，帮助企业实现智能决策，提高推荐准确度，进而优化客户体验，实现定制化服务，使产品、服务的营销方案更符合市场需求。

6. 语音识别

语音识别技术就是让智能设备听懂人类的语音，涉及数字信号处理、人工智能、语言学、数理统计学、声学、情感学及心理学等多个学科。这项技术可以提供比如自动客服、自动语音翻译、命令控制、语音验证码等多项应用。随着人工智能的兴起，语音识别技术在理论和应用方面都取得大突破主要应用领域包括语音识别听写器、语音寻呼和答疑平台、自主广告平台，智能客服等。语音助手是人工智能语音识别的典型应用之一，如苹果的 Siri、亚马逊的 Alexa、谷歌的 Google Assistant 等，在手机、智能音箱等设备上被广泛使用；智能音箱通过语音识别技术，实现与用户的交互。另外，许多公司在客服中引入语音识别技术，可以通过自动语音识别和自然语言处理技术，实现自动化的客户服务。语音翻译技术可以将语音实时转换为文字，并进行翻译，实现跨语言交流。人工智能语音识别技术在准确性和稳定性上已经取得了显著进步，但仍面临一些挑战，如环境中的噪声干扰、不标准的语音识别率、容错率的处理等问题。随着技术的不断创新和发展，

人工智能语音识别将在更多领域得到应用并不断提升用户体验。未来我国的智能语音识别技术将会随着科技的发展逐渐进步，最终应用到生活的方方面面。

7. 图像识别

图像识别是指利用计算机对图像进行处理、分析和理解，以识别各种不同模式的目标和对象的技术，是应用深度学习算法的一种实践应用，应用场景，包括电子商务，游戏，汽车，制造业和教育。现阶段图像识别技术一般分为人脸识别与商品识别，人脸识别主要运用在安全检查、身份核验与移动支付中；商品识别主要运用在商品流通过程中，特别是无人货架、智能零售柜等无人零售领域。成熟的图像识别技术加以人工智能的支持可以自行对视频进行分析和判断，发现异常情况直接报警，带来了更高的效率和准确度；在反恐领域，借助机器的人脸识别技术远远优于人的主观判断。

许多科技巨头也开始了在图像识别和人工智能领域的布局，Facebook 签下的人工智能专家 Yann LeCun 最重大的成就就是在图像识别领域，其提出的 LeNet 为代表的卷积神经网络，在应用到各种不同的图像识别任务时都取得了不错效果，被认为是通用图像识别系统的代表之一；Google 借助模拟神经网络“DistBelief”通过对数百万份 YouTube 视频的学习自行掌握了猫的关键特征，这是机器在没有人帮助的情况下自己读懂了猫的概念。

3.2.2 行业应用

1. 自动驾驶

自动驾驶,是依靠计算机与人工智能技术在没有人为操纵的情况下,完成完整、安全、有效的驾驶的一项前沿科技。从技术角度来看,人工智能已经可以实现自动驾驶车辆在复杂环境下的感知、决策和控制,通过深度学习、机器视觉和传感器融合等技术手段能够对道路、交通标志和其他车辆进行实时感知和识别,快速做出决策并控制车辆动作。从行业应用角度来看,人工智能在自动驾驶领域已经得到广泛应用。在汽车行业,人工智能能够帮助车辆实现自动驾驶功能,提高行车安全和驾驶舒适性;在物流行业,人工智能能够实现自动驾驶的无人配送车辆,提高配送效率和降低成本;在城市交通管理领域,人工智能能够实现车辆的协同驾驶和交通流量优化,改善交通拥堵和环境污染问题。我国相关的信息技术、人工智能技术、汽车制造技术等与世界先进水平仍存在一定的差距,自动驾驶技术相对不够完善,其实际的应用情况存在较大的质量波动,需要进一步进行技术研发的投入,提升汽车自动驾驶的安全性、效率性和可靠性。

2.金融风险评估

神经网络、专家系统、支持向量机以及混合智能等人工智能模型在金融风险评估领域的应用能够提高数据处理速度、加深数据分析深度、降低人工成本,从而提升金融风险控制的效能。近年来无论是传统金融机构、消费金融机构还是互联网金融公司,都在加紧智能化系统建设或者对外合作,实现智能化风控。人工智能可以实现对风险及时有效的识别、预警与防范,智能反欺诈能够通过分析大量的数据,识别出欺诈行为。机器学习可以通过对历史欺诈案例的分析,识别出

欺诈的模式和规律。当新的欺诈行为出现时，机器学习可以通过对这些模式和规律的匹配，自动识别出欺诈行为。目前，智能风控面临的挑战主要包括数据的全面性、真实性及数据挖掘效率等。智能风控目前还是比较依赖大数据还有专家规则，只有在正确数据基础之上才能正结论，当数据量很大时，数据真实性及是否数据被污染，无法进行有效鉴别。

3. 量化交易

量化交易是指以先进的数学模型替代人为的主观判断，利用计算机技术从庞大的历史数据中海选能带来超额收益的多种“大概率”事件以制定策略，进而交易的过程，其核心是用数学模型或者明确的交易规则指导交易，而不是纯主观判断。人工智能可以对海量的市场数据进行分析 and 预测，识别出潜在的市场趋势和机会。通过机器学习算法，它可以自动学习和优化交易策略，根据预测结果和设定的交易策略，自动化地进行交易决策，并根据市场情况和指标，动态调整交易策略，实现自动化下单和止损，可以在毫秒级别的时间尺度上进行高频交易，通过快速分析市场数据和执行交易，实现瞬间的套利和利润。另外，人工智能可以帮助量化交易者进行风险管理，通过对交易风险的预测和监控，提供实时的风险警示和建议。人工智能和量化交易的融合发展推动了智能投顾、量化基金、智能金融等众多领域的发展，如中国银行、中国农业银行、浦发银行、华夏基金等金融机构均推出了自己的智能投顾服务。

3.3 智能算力发展挑战及解决方案

3.3.1 算力需求

首先，算力需求急剧上升。根据 IDC 的预测，未来 3 年全球新增的数据量将超过过去 30 年的总和，到 2024 年，全球数据总量将以 26% 的年均复合增长率增长到 142.6ZB。这些将使得数据存储、数据传输、数据处理的需求呈现指数级增长，不断提升对算力资源的需求。另外，面向人工智能等场景，大规模的模型训练和推理也需要强大的高性能算力供应。其次，算力灵活调度受限。另外，不同应用场景对 AI 算力的精度、能效、速度、交互性、部署方式以及网络安全、网络带宽的要求各不相同，部分场景难以通过网络实现算力的灵活高效调度，无法满足人工智能推理和训练需求。

解决方案：一是为满足业务的大算力需求，通过构建智算中心、超级计算中心以及云计算中心实现对大算力业务的资源供给，另外还可以通过网络将数据源周围闲散算力(云计算、边缘计算等)调度起来以弥补大型科学装置的算力缺口。二是提高算力使用效率，如算网协同优化 AI 计算效率，算力服务结合人工智能技术推动算力资源的精准配置和按需获取。三是发展可扩展、自演化、高可靠和安全的新型网络架构，促进数据的处理与流通。四是建立统一的算力调度平台，形成覆盖全国、互联互通的算力调度服务体系 and 平台基础框架，实现对全网算力资源统一编排、统一输入输出。

3.3.2 能耗

人工智能算法高算力需求导致大量能源消耗。首先是芯片能耗，人工智能训练和推理需要大量的计算资源，随着 AI 算力的逐步提升，

能耗和成本也在逐渐增加。高性能计算设备在运行过程中会产生大量热量,芯片是集成超大规模电路,随着晶体管密度和时钟频率的提高,芯片功耗大幅增加,同时,电源电压和阈值电压的降低导致漏电流增加。**其次是系统级能耗**,功耗过高会导致芯片温度升高,需要进行散热和冷却,为了维持设备的温度在可承受范围内,需要消耗额外的能源来驱动冷却系统。

解决方案:一是针对算力基础设施风火水电的节能。首先,算力基础设施的建设前优先考虑算力中心的地理位置,例如亚马逊、谷歌倾向于将数据中心建在天气寒冷的爱尔兰,微软将数据中心建在海里,依靠海水的温度来为数据中心降温。在算力基础设施建设后,推动数据中心采用液冷技术来满足服务器大功耗高密度部署带来的散热需求,液冷技术具有比热容大、散热效率高、降低能耗等优点。目前浸没、喷淋、冷板等液冷主要部署方式都已有市场应用。二是针对业务层面的合理安排调度,寻找在时间上相互匹配的业务。例如优先处理用户驱动型业务,并在计算资源闲时处理结果驱动型业务,充分发挥算力资源的能力,以提高资源使用效率。另外在技术层面使用弹性扩容等技术,在计算资源闲时,关闭部分服务资源以节约能耗。

3.3.3 算法复杂度

智能算力应用场景复杂化,数据量及算法的复杂度急剧增加,相应程序运行所需要时间和空间(内存)资源不断攀升。计算规模从单机到集群再到大规模云计算,规模呈量级、指数级增长,计算架构从单一通用架构 CPU+GPU 到混合异构架构 CPU+GPU+FPGA+xPU,系

统环境方面从单一用户以及单一场景解决到现在多个用户、多场景复杂环境的构建，数据规模和模型复杂度的增加导致时间复杂度可能变得非常高，出现计算时间长而效率低下的问题。一些复杂的算法在执行过程中需要大量的内存空间来存储数据和中间结果，而设备或平台的资源受限，导致在计算过程中高效地存储和访问数据成为巨大挑战。

解决方案：一是积极进行算法优化研究，通过优化算法的设计和实现，减少算法的时间复杂度和空间复杂度，可以使用更高效的数据结构、减少循环次数、适当使用缓存等技术来提高算法性能，采用并行计算、分布式计算等技术加速算法的执行。二是数据压缩和存储优化，对于算法中处理的数据，可以使用压缩算法进行压缩，减少数据的存储空间和传输时间。同时，优化数据的存储结构，选择合适的数据类型和存储方式，减少内存占用和访问时间。三是利用机器学习和自动化等人工智能技术，通过对算法执行过程进行学习和调优，提高算法的执行效率。可以使用机器学习算法来寻找最优的算法参数，自动调整算法的执行策略，提供更高效率的算法解决方案。

3.3.4 数据隐私和安全

人工智能充分挖掘数据价值，数据泄露和滥用成隐患。智能算力的发展使得大规模数据的收集和处理变得容易，如何在保护用户隐私的前提下，充分利用数据进行算法训练和应用，是需要解决的难题。首先是人工智能系统需要大量的数据来进行训练和学习，这些数据可能包含个人身份信息、偏好等敏感信息，如果这些数据泄露，可能会导致个人隐私暴露和身份盗窃等问题。另外，人工智能系统通过分析

大量数据来预测和识别个人的行为模式和习惯，但这种大规模监控是否是对个人隐私权的侵犯需要认真评估。人工智能系统采集的数据可能被恶意利用来进行广告定位、骚扰、诈骗等，若人工智能系统存在安全漏洞，黑客可以利用这些漏洞来入侵、篡改或者破坏系统，造成重大安全事故。

解决方案：一是构建以数据为中心的数据安全治理平台。具体包括建立数据安全风险感知体系，实现对数据的态势可知、威胁可现、风险可控；打造零信任的数据安全机制，通过身份认证、权限管理、安全审批、安全审计、安全感知和安全策略控制打造完整的零信任安全机制；建设数据质量管理智能平台，建立数据质量核验任务，自动完成数据质量规范性、一致性、准确性和完整性的检查。二是加快标准研制和试点推广工作。研制人工智能安全参考框架、数据集安全、数据标注安全、开源框架安全、应用安全和服务能力要求等标准；同步开展人工智能基础性标准研究工作，研究应用安全风险评估类标准及智能制造、智能网联汽车等重点人工智能产品和服务类安全标准，逐步推进其他领域人工智能安全标准的研究工作。

3.3.5 生态合作

智能算力的持续发展需要跨领域的合作和资源共享，面临着交互标准与互操作性、供应链管理以及盈利模式与利益分配等挑战。AI、物联网、大数据等新兴技术的注入将促使细分场景的应用以指数级增长，智能算力的应用涉及多个领域，包括人工智能研究、硬件开发、数据科学等。智能算力的持续发展需要跨领域的合作和资源共享，各

个领域的标准和技术往往不统一，缺乏互操作性。另外智能算力产业涉及多个环节和参与方，包括算力提供方、应用开发方、数据提供方等，如何有效管理供应链，保证各方的合作效率和利益平衡是一大问题。要想把算力真正转化为生产力，还需要多重因素支撑，如今仍面临使用门槛有待提高、融合技术有待突破、场景应用有待优化等问题。

解决方案：一是上下游建立合作伙伴关系，创新合作模式。与相关领域的企业、机构建立合作伙伴关系，共同推动算力的发展，可以通过合作共享资源，实现互利共赢的合作关系，例如建立联合研发中心、共享经济平台、组建产业联盟等。二是构建开放的计算平台，为上下游合作伙伴提供统一的接口和标准，方便资源对接和共享。三是制定统一的行业标准，规范算力提供方和使用方之间的合作流程和要求。四是政府出台相应的政策和规划，促进算力上下游合作的发展。给予相关领域企业和项目支持、减税等政策优惠，鼓励企业进行技术创新和合作，推动算力生态的健康发展。

4 智能算力未来发展趋势

4.1 人工智能加速渗透，多样化场景催生多元化算力需求

人工智能在各行各业的应用离不开海量数据的处理、存储和云化，伴随着 5G 商业化进程加速、流量持续增长，云计算和边缘计算需求会继续增加，从行业趋势和应用需求看，多样性计算时代正在到来。IDC 预测在数字智能创新阶段，数字化普及率上升到新的高度，应用规模将从百万级上升到千万级，连接数上升到百亿级，智能算力将成为基础能力，这一阶段的显著特点使传统单一架构难以满足要求，对

计算平台提出了新的挑战，驱动计算架构向多样性发展。源于多种数据类型和场景驱使计算架构的优化，多种计算架构的组合是实现最优性能计算的必然选择，其中，边缘侧需要 AI 算力，数据中心侧要处理和存储海量数据，需要高并发、高性能，特别是高吞吐的算力。预计在未来 10 年内，智能算力将向多元化发展，提供更高的计算性能。

4.2 政策驱动，智能算力低碳发展成硬性要求

传统的计算和数据处理方式往往需要大量的能源消耗，并且会产生大量的碳排放，而采用低碳算力可以显著减少碳排放量，降低对环境的影响。马萨诸塞大学阿默斯特校区的研究人员最新的论文结果表明，训练一个 AI 模型产生的能耗多达五辆汽车一生排放的碳总量。

《贯彻落实碳达峰碳中和目标要求推动数据中心和 5G 等新型基础设施绿色高质量发展实施方案》，以及上海市、甘肃省、云南省、江苏省等地出台的数据中心相关政策文件，都对数据中心绿色低碳提出明确要求。算力基础设施碳排放相关标准已逐步制定和发布，四部门明确提出数据中心绿色低碳达到 4A 等级，数据中心低碳等级评估中三大运营商、第三方数据中心企业和科技企业的 20 余个数据中心已经通过测试评估，绿色低碳发展已成为各级政府的关注焦点和建设推进的基本要求，面对全球对于减少碳排放和应对气候变化的呼吁，低碳算力将成为我国算力发展的重要方向。

4.3 边缘智能应运而生，边缘计算与人工智能融合发展

边缘计算将计算资源从云中心转移到网络边缘侧的服务器，为联网的终端设备提供计算支持，通过将算力传递到设备和传感器端，可

以更快速地进行实时处理和决策，减少对网络的依赖，同时保护数据隐私。与此同时，人工智能迅猛发展，在工业物联网、自动驾驶、智能家居等领域得到了广泛地应用，为人们的生产和生活带来了极大的便利。在边缘智能中，边缘计算和人工智能二者相互受益，边缘计算利用人工智能对边端进行智能地维护和管理，人工智能在边缘计算平台上提供智能化的服务，通过在边缘节点上进行数据的计算和分析，减少数据传输和处理的延迟，提升智能应用的实时性。未来边缘智能将在公共安全、智能交通、智能制造、智能驾驶等诸多场景得到广泛应用，大量的智能设备被部署在边缘节点上，边缘侧将成为整个网络数据汇聚处理的最前线，如何应对海量异构数据的冲击是边缘计算技术的重要挑战，数据的筛分、整合、存储、访问、安全管理等也将成为边缘智能的技术研究热点。

4.4 智算中心建设加速，应对高质量算力需求

智算中心以异构计算资源为核心，通常面向人工智能训练和推理的需求，因其专用性，在面向人工智能场景时性能和能耗更优，借助人工智能芯片+算力机组的强强组合，算力可以实现指数级别的提升。另外，智算中心有利于提高算力安全可用性，从算力卡到服务器自主打造整个算力“底座”的核心部件不仅针对性更强、效率更高，还更加自主可控、安全可靠，更能确保智算中心安全稳定运行。智算中心从早期实验探索逐步走向商业试点，尽管现有规模占比不高，但随着我国各类人工智能应用场景的丰富，智算需求将快速增长，预期规模增速迅速爆发。未来的智算中心建设将采用多元开放的架构，兼容成熟

主流的软件生态，支持主流的 AI 框架、算法模型、数据处理技术以及广泛的行业应用。

4.5 模型规模不断扩展，海量多元化数据亟需巨量化算力

通过大规模数据训练超大参数量的巨量模型，是实现通用人工智能的一个重要方向。模型规模的扩展可以提供更大的计算资源和存储能力，帮助算法实现更复杂的学习和推理过程，并提高算法的通用性和迁移能力。随着人工智能应用范围不断扩大，未来 10 年内，预计人工智能模型的规模将进一步扩展。目前，谷歌、微软、英伟达、新华三、智源人工智能研究院、百度、阿里等全球知名的 AI 公司相继推出了各自的巨量模型，未来将有更多的超大规模模型出现，拥有数十亿个参数。目前大模型主要集中在自然语言处理领域，多模态任务领域也有一定突破，随着大模型基础设施和垂直行业领域小模型应用的发展，围绕上中下游将产生丰富的大模型产业链，大模型将更广泛的赋能各行各业应用。面向未来产业界更为复杂的智能决策场景，基于多种网络数据预训练，具有决策能力的大模型也将是下一步发展的重点。大模型加速社会各领域数字化转型及智能化发展，未来将需要更强大的算力来进行训练和推理，以应对更复杂的人工智能任务。

4.6 自主学习能力提升，推动算力实现更高层次智能

随着人工智能技术的不断进步和应用需求的增加，算法具备自主学习的能力将成为推动智能算力发展的关键因素。传统的算法通常需要人工设计和调整，适用于特定的任务和数据，而具备自主学习能力的算法能够通过分析和处理数据，自动调整模型参数，优化学习策略，

从而适应不同的任务和数据场景，提高算法的适应性和灵活性。通过具备自主学习能力的算法，智能系统可以从海量的数据中提取有价值的信息和知识，实现自主的决策、学习和创新。另外，目前的人工智能模型通常需要大量的标注数据进行训练，耗费大量人力和时间成本，而通过强化学习、迁移学习等技术的发展，模型可以更好地从少量数据中提取知识，并进行迁移和泛化，从而减少对大规模标注数据的依赖。

5 智能算力发展展望与建议

5.1 智能算力发展展望

智能算力需求爆发。当前人工智能技术正加快融入千行百业，超大规模人工智能模型和海量数据对算力的需求也持续攀升。诸如云游戏、元宇宙、VR/AR 等新应用场景加速照进现实，算力需求旺盛，而大模型的训练和推理过程进一步带动算力需求爆发，同时也推动算力需求由通用性 CPU 算力向高性能 GPU 算力发展。根据 IDC 预测，中国智能算力规模将持续高速增长，预计到 2026 年中国智能算力规模将达到 1271.4EFLOPS，未来五年复合增长率达 52.3%，同期通用算力规模的复合增长率为 18.5%。

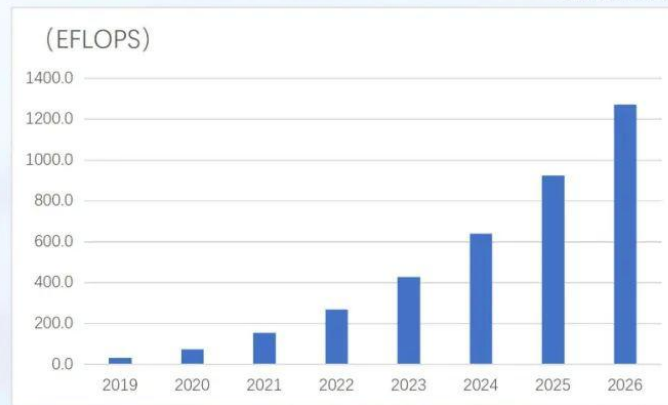


图 10 我国智能算力规模及预测

(数据来源: IDC, 2022)

智能算力赋能千行百业。随着人工智能技术的持续突破,智能时代将加速到来。从技术上来看,人工智能与其他数字技术将会有更广泛融合、碰撞,如“AI+量子计算”、“AI+区块链”、“AI+5G”、“AI+AR/VR”,拓宽人工智能应用场景。从应用领域看,人工智能将赋能各行各业,广泛获得更加多元的应用场景和更大规模的受众。AI已经渗透到工业、医疗、智慧城市等各个领域,未来会有更多产业与智能技术进行创新融合,催生出更多新业态、新模式。从支撑能力看,依托坚实的智能算力支撑,人工智能技术将逐渐转变为像网络、电力一样的基础服务设施,向全行业、全领域提供通用的AI能力,为产业转型打造智慧底座,促进产业数字化升级和变革,生物医药、天文地理等科技领域将产生一大批新的研究成果,智能驾驶、影视渲染水平大幅提升,广大人民群众在日常生活中能够切身体会到算力带来的变化。

5.2 智能算力高质量发展建议

随着新一轮科技革命和产业变革深入发展,人工智能渗透到各行各业,智能算力在赋能产业发展、促进数实融合方面将发挥更加显著的作用,其带动产业创新的“乘数效应”也将进一步放大,在未来数年内将为各领域创新发展注入新的活力。目前,智能算力还面临诸多挑战,一方面,海量数据呈指数级增长,数据在加速流动;另一方面,高算力需求还存在能耗高、算法越来越复杂、数据隐私和安全性等多方面的挑战和问题,亟待加以重视,出台政策推动相关产业发展。

产业方面,在国家战略层面制定规划,多举措推动智能算力健康有序发展。**第一,**加强我国人工智能芯片的研制。制定中国的人工智能芯片国家发展战略,系统推进人工智能芯片产业发展,产学研用联动,推动成果转化,形成人工智能芯片良好的产业生态。**第二,**构建统一的人工智能算力服务中心和孵化平台。解决算力昂贵、算法软硬不解耦、传统应用场景算法众多、选择难、新应用场景算法缺乏生态支撑等问题。**第三,**加强资金支持。引导国家对基金和市场化的社会资本有序、持续地进入,国家资本市场监督管理部门有倾向性地对智算相关企业予以更多政策倾斜,构建持久、有序、宽容的资本环境。**第四,**多措并举推进绿色智能算力发展。加强节能降碳技术创新应用,推动液冷、蓄冷、高压直流、余热利用、蓄能电站等技术在算力基础设施建设中推广应用,同步提升太阳能、风能等可再生能源利用水平。

技术方面,加大对智能算力领域的技术研发投入,加大创新攻关。将计算机视觉、自然语言处理、机器学习等各类智能算力技术进行整

合，实现多模态的算力开发，提供更丰富的智能应用。加快推进软硬件适配，提高计算效率和资源利用率，针对不同的应用场景，研发更优化的算法，提供更多高性能、高效能的算力解决方案，鼓励跨领域的技术创新发展可扩展、自演化、高可靠和安全的新型网络架构，通过新型网络架构实现数据资源与算力需求的敏捷对接和智能匹配。构建从智能芯片到算法框架，从行业大模型到应用的全站式产业链，加快人工智能发展。

标准方面，加快推动开放标准建设，将多元化算力转变为可调度的资源。促进各部门间的协同合作，共同制定智能算力的技术标准和规范，推动行业的规范化和标准化发展。建立和完善智能算力数据安全的标准体系，保障用户数据的隐私和安全，加强对个人信息的保护。建立公平、公正的竞争机制，推动产业链中各环节的公平竞争，促进整个智能算力产业的健康发展。