IDC 浪潮信息

2022-2023
中国人工智能计算力发展评估报告

目录

核心观点

1*

算力是数字经济时代的核心生产力，智算力则是数字化创新的源动力。人工智能已在国家经济建设、科技实力提升、推动生产力发展等方面呈现出举足轻重的作用。人工智能算力正在为国家创造力的发展带来实质性推进，为企业带来切实的创新成效，加速驱动新业态的形成。

2*

我国进一步明确了人工智能对于提升国家核心竞争力的重要支撑作用，随着新基建、数字经济等持续利好政策的推动，中国人工智能市场将保持平稳增长。IDC预测，2022年中国人工智能市场相关支出将达到130.3亿美元，有望在2026年达到266.9亿美元，2022至2026年年复合增长率达19.6%。

3*

中国算力规模，尤其是智能算力规模，正在高速增长。2021年中国智能算力规模达155.2每秒百亿亿次浮点运算（EFLOPS），2022年智能算力规模将达到268.0EFLOPS，预计到2026年智能算力规模将进入每秒十万亿亿次浮点计算（ZFLOPS）级别，达到1,271.4EFLOPS。2021年中国通用算力规模达47.7EFLOPS，预计到2026年通用算力规模将达到111.3EFLOPS。2021-2026年期间，预计中国智能算力规模年复合增长率为52.3%，同期通用算力规模的年复合增长率为18.5%。伴随人工智能算力需求的高速增长，建立健全助商惠民的数字基础设施服务体系、推进算力基建化发展势在必行。

4*

人工智能服务器仍是人工智能市场增长的主力军。IDC数据显示，2021年全球人工智能服务器市场的同比增速为39.1%，超过全球整体人工智能市场增速（20.9%），是整体人工智能市场增长的推动力。在中国，人工智能应用的加速落地很大程度推动了中国人工智能服务器市场的高速增长。2021年人工智能服务器市场规模59.2亿美元，与2020年相比增长68.2%，预计到2026年，中国人工智能服务器市场将达到123.4亿美元。

5*

目前，中国仍以GPU为主实现数据中心计算加速，市场占有率近90%，但ASIC、FPGA、NPU等非GPU芯片市场也在加速发展。近年来，在政策、资本等多重因素驱动之下，中国人工智能芯片专利数量不断增长，产业链和应用场景不断完善扩充，市场规模不断扩大，产业正向好发展。

6*

通过aaS服务提供AI平台和AI服务正越来越被用户接受，AI平台和服务的快速迭代能力和丰富的场景化人工智能能力，正在为行业的智能化发展提供有力支撑。IDC调研显示，目前排名前三的人工智能云服务是：搜索、人脸识别和推荐引擎，预计未来18个月，排名前三的人工智能云服务将为：自然语言处理、图像识别和视频识别。

7*

人工智能算法模型呈现出多样化、巨量化、专业化等显著特征，算法基建化发展对于实现普惠人工智能具有重要作用，绿色高效、可应用性强等成为主要诉求。市场积极探索面向专业场景的轻量化模型以加速落地运作，并通过集中式的数据和算力开发模式为企业提供预训练平台，提供分布式加速计算集群解决方案，合理匹配计算任务与计算资源，提升整体利用率和训练效率，加速实现人工智能普惠化目标。

8*

从计算架构发展来看，基于 DSA（Domain-Specific Architectures）思想设计的人工智能芯片正在成为主导，推动了人工智能芯片多元化发展。多元算力从“能用”到“好用”并且为企业创造业务价值，离不开通用性强、绿色高效、安全可靠的计算系统的支持。业内正在推动多元算力系统架构创新，基于计算节点内和节点间的互联技术破局现有计算架构的瓶颈，通过充分调动起多芯片、多板卡、多节点的系统级能力，实现各种加速单元以及跨节点系统的高效协同，提升计算性能。

9*

2022年，人工智能在各个行业的渗透度均有提升，应用渗透度排名前五的行业依次为：互联网、金融、政府、电信和制造。总体来看，人工智能在各个行业的应用程度都呈现不断加深的趋势，应用场景也越来越广泛，人工智能已经成为企业寻求新的业务增长点、提升用户体验、保持核心竞争力的重要能力。

10*

在2022年中国人工智能城市排行榜中，北京、杭州、深圳继续保持前三名，上海和广州分列第四、五名，天津进入前十名。除了 TOP10 城市之外，诸如合肥、武汉、长沙等多个城市在自身产业优势及各种因素推动下，人工智能应用取得了较大进展，未来将会出现越来越多具有城市特点的人工智能示范区，为产业发展树立标杆。

人工智能发展概况

1

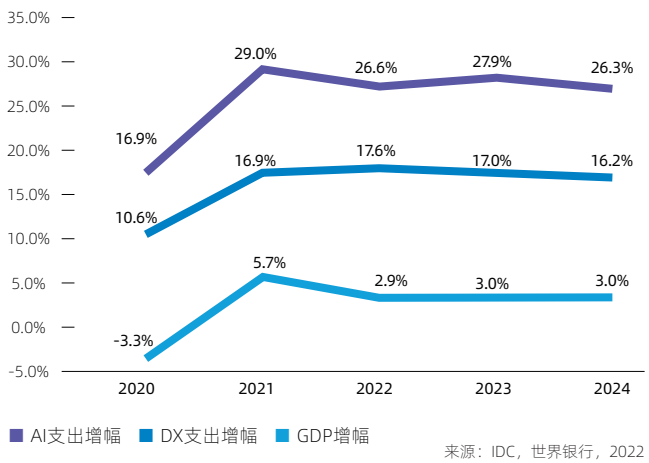
1.1 放眼全球：各国持续加大布局，拓展人工智能创新应用

■ 全球企业持续增加对人工智能的投资

伴随数字经济的持续发展，以及诸如新冠肺炎疫情等突发事件的影响，企业积极打造敏捷反应机制，推进精益化管理，提升组织创新能力，以期在变革中发现新的赛道甚至实现弯道超车，实现数字时代中的自身发展。

为满足企业内部发展需求和外部市场需求，企业一直大力投资数字化转型相关技术，特别是在人工智能领域。人工智能支出已经成为支持企业数字化转型支出的主力之一。IDC数据统计，全球范围内，企业在包括硬件、软件和服务在内的人工智能（AI）市场的技术投资从2019年的612.4亿美元增长至2021年的924.0亿美元，预计将在2022年（同比）增长 26.6%至1,170.0亿美元，并有望到2025年突破2,000亿美元，增幅高于企业数字化转型（DX）支出整体增幅。

图1 全球人工智能支出、数字化转型支出及GDP增长趋势预测，2020-2024



来源：IDC，世界银行，2022

■ 全球各国人工智能发展特征

对于国家而言，人工智能在国家经济建设、科技实力提升、社会生产力发展等方面表现出举足轻重的作用。世界诸多国家，尤其是领先的大国正在竞相开发和部署人工智能技术，以改善人民生活、工作、科研条件和方式，进而在未来智能世界中保持科技竞争力和优势，同时对经济发展带来促进作用。

纵观各国，人工智能发展呈现如下特征：

- **各国持续布局人工智能发展战略：**诸如美国、英国、欧盟、日本等全球主要经济体进一步推进人工智能战略的制定，在国家战略领域的引导下，持续推进技术的产业化发展：在2022财年美国研发类项目预算中，人工智能、量子计算等领域被列为优先投资项目；英国采取全行业、大投入策略，从基础设施建设到人工智能企业进行全方位扶持建设，一批人工智能创业企业崭露头角，得到市场认可；欧盟、日本等尝试结合产业优势，成为细分领域“领头羊”，比如，德国结合工业行业积累，积极打造人工智能工业制造品牌；日本从可持续发展、灾难预警治理等方面入手，希望构建具有国家特色的人工智能产业。在政府推动下，各国科技企业积极跟进，进一步提升人工智能领域投资。这些人工智能发展战略对应用水平的提高起到了促进作用，同时也推动了各国对人工智能的投入，根据IDC数据，各个国家的人工智能投入在其本国GDP总量的占比均出现不同程度提升，从2015年-2022年，美国人工智能投入占比提高了3倍，德国提高了5倍，而中国人工智能投入占比则提高了13倍之多。全球领先超大规模企业引领了这一发展，微软支出数十亿美元投资AI创业公司OpenAI构建智能计算系统；Stability AI公司在 AWS 云上部署了4,000多个节点的集群用于人工智能训练工作负载等。

- **加强政策支持和引导，探索监管与创新之间的平衡：**各国越来越重视公共政策的出台，以期规范和引导人工智能的良性发展：诸多国家和地区相继出台不同的人工智能监管法规，例如，美国国防部2022年出台的《负责任的人工智能策略》报告，指出政府应主动思考并提出应对策略，解决人工智能可能带来的数据安全、道德伦理等问题；日本政府发布的《人工智能战略2022摘要》指出产业发展对数字化、人工智能等基础设施具有迫切需求，在发展的前提下，政府也会配套关注人工智能安全策略；2022年，欧洲议会决议通过《欧洲机器人技术民事法律规则》，填补机器人和人工智能民事立法空白，加强对机器人市场的监管和引导。各国一方面鼓励区域性的人工智能解决方案的提出，降低在数据获取、数据偏见、隐私和道德伦理等方面存在的潜在风险，另一方面也在思索如何在监管、创新和人工智能商业化之间取得平衡。

- **发挥人工智能在社会与民生问题方面的价值：**各国政府面临的重要民生问题也有望通过人工智能技术制定更优的预防和管理方案。在医药研发领域，人工智能模型对基因组结构预测提供数万倍计算效率提升，同时精度并未有明显损失，有效助力新冠疫苗研发；在公共卫生事务管理领域，基于人脸识别的自动健康码扫描、身份证识别等服务，为突发公共事件的治理和预防提供新的有效方案；在

气象灾害预测领域，越来越多组织以神经网络为基础进行智能化风险识别，比如，日本将“利用人工智能应对紧迫危机”作为2022年首要人工智能战略，希望通过利用人工智能技术，实现大规模灾害的精准预测，提升国家灾害抗压力；在农业领域，各国也加强人工智能赋能，以美国为例，美国国家科学基金会（NFS）联合各政府机构以及科技巨头，共同合作新建国家人工智能研究所，以期提供可持续发展的作物生产解决方案，对抗日益凸显的粮食危机。

人工智能的快速发展，极大拉动了算力的发展。如同生物大脑是“人智”的核心，人工智能也同样非常依赖一个高质量的“大脑”，即人工智能基础设施，包含计算、存储和网络。面对数据攀升、算法和模型领域的突破，“大脑”需要尽可能快速、精准地处理大量数据或执行复杂的指令，这对人工智能计算力提出更高的需求。

当前，全球人工智能算力基础设施产业加速发展，为人工智能技术更广泛的场景落地带来可能。释放算力的价值，对国家整体的经济发展将起到推动作用。根据《2021-2022全球计算力指数评估报告》的研究结果，计算力指数平均每提高1点，数字经济和GDP将分别增长3.5‰和1.8‰。国家计算力指数越高，对经济的拉动作用越强。

■ 元宇宙迎来快速发展风口期，算力基础设施重要性凸显

2021年元宇宙（Metaverse）成为全球科技领域备受瞩目的重要概念之一，它作为混合现实类先进解决方案可推进实现数实融合：一方面实现虚拟数字世界对现实物理世界的映射，另一方面也在虚拟数字世界里创造可以与现实世界交互的新体验，进而打造共情化的卓越体验，实现高效有序的韧性运营。

在这个过程中，人工智能技术将对元宇宙的建设起到至关重要的作用，计算机视觉、机器学习、自然语言处理等人工智能技术为创造广阔复杂的虚拟空间、优化交互体验提供底层支持，元宇宙的实现将营造一个全新的数字环境。在这个环境里，诸如3D场景构建、实时渲染、高仿真交互等场景的实现需要大规模算力支持。当下，尽管全球各大互联网巨头已加入元宇宙战略布局，但元宇宙的发展仍处于萌芽阶段，打造趋于终极形态的元宇宙世界需要大量数字化内容的创建和分发、资产的数字化建设以及可落地场景的探索，这对算力基础设施将提出更高的要求，也在计算成本等维度给企业带来更多的挑战。

目前元宇宙对云计算、边缘计算、人工智能芯片等领域的上下游生态圈已经产生推动作用，IDC预计在2022年，以运营商、基础架构供应商、互联网大厂为主的元宇宙基础设施建设参与者会领衔发力，加快相关产品的研发、操作平台的创新和内容的升级，除在游戏、媒体和娱乐等行业内的诸多场景外，培训和远程办公有望成为最先落地的元宇宙商用场景。此外，元宇宙也将在教育、金融、零售、文化旅游、智慧城市等领域实现渗透，为人工智能的发展带来更多创新空间。企业应提前布局底层算力支持平台，为更多元宇宙场景的落地和运行提供算力支持。

1.2 聚焦中国：人工智能算力需求稳增，持续夯实算力底座

■ 政策驱动下，中国人工智能发展迎来黄金时期

中国始终强调科技兴国的重要性。数字经济时代，技术的力量更为凸显。近年来，中国政府相关部门相继发布一系列政策，更加明确了人工智能对于提升中国核心竞争力的重要支撑作用，加上新基建、数字经济等持续利好政策的推动，中国人工智能市场保持平稳增长。IDC预测，2022年中国人工智能市场相关支出将达到130.3亿美元，有望在2026年达到266.9亿美元，2022至2026年年复合增长率达19.6%。

加速技术行业落地、推进优化治理是中国人工智能相关政策的核心目标方向：

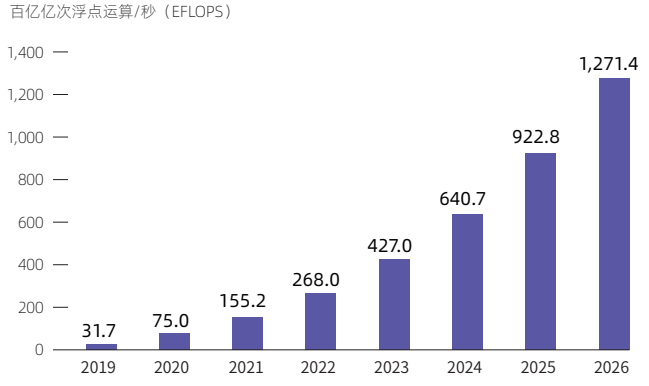
- **加速技术的行业落地：**《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》（简称“十四五”纲要）指出要推动互联网、大数据、人工智能等同各产业的深度融合。基于此，各省市纷纷布局打造示范应用行业和场景，推进人工智能与产业融合。2022年《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》（简称《意见》）发布，《意见》为各地方和各主体加速人工智能行业和场景化落地提供指引，指出制造、农业、物流、金融、商务、家居等重点行业可深入挖掘，“促进智能经济高端高效发展，以更智能的城市、更贴心的社会为导向，在城市管理、交通治理、生态环保、医疗健康、教育、养老等领域持续挖掘人工智能应用场景机会，开展智能社会场景应用示范”，并鼓励科研机构及高等院校加大对于人工智能技术研究、开发及应用的投入力度。

- **推进优化治理：**伴随人工智能应用逐步广泛落地，国家相关机构通过制定一系列规范和政策，例如要求医疗人工智能影像公司“持证上岗”，制定智能网联汽车道路测试与示范应用，完善无人驾驶领域相关监管框架和法规等，旨在一方面监管和把控人工智能发展过程中存在的潜在风险和问题，另一方面基于规范化管理促进人工智能的规模化、可持续化发展。此外，数据安全及个人隐私也成为新一代人工智能治理的重要组成部分，基于联邦学习、隐私计算等技术创新，法律法规约束和社会监督，旨在降低安全隐患，避免乱用滥用，促进企业重新审视自身数据及人工智能战略，调整和优化管理流程，以符合政策法规，促进科技向善。

■中国智能算力规模持续扩大，推进算力、算法基建化发展势在必行

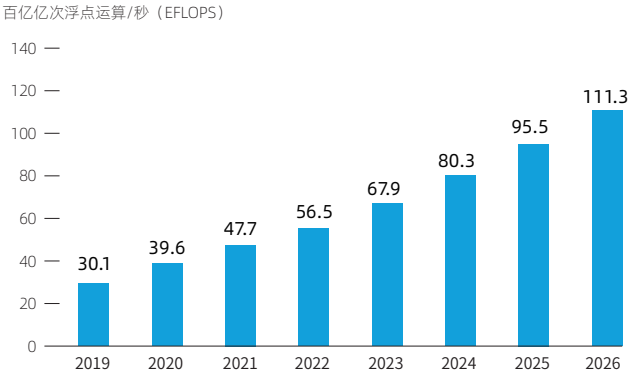
数据海量增加，算法模型愈加复杂，应用场景的深入和发展，带来了对算力需求的快速提升。为定量评估算力规模的大小，本报告基于《IDC中国加速计算服务器市场半年度跟踪报告》及智能加速卡半精度（FP16）相当运算能力数据，测算了中国智能算力规模。结果显示，中国智能算力规模正在高速增长。2021年中国智能算力规模达155.2 每秒百亿亿次浮点运算（EFLOPS），2022年智能算力规模将达到268.0 EFLOPS，预计到2026年智能算力规模将进入每秒十万亿亿次浮点计算（ZFLOPS）级别，达到1,271.4EFLOPS。作为参考，本报告基于《IDC中国服务器市场季度跟踪报告》及CPU双精度（FP64）运算能力数据，测算了中国通用算力规模。2021年中国通用算力规模达47.7EFLOPS，预计到2026年通用算力规模将达到111.3 EFLOPS。2021-2026年期间，预计中国智能算力规模年复合增长率达52.3%，同期通用算力规模年复合增长率为18.5%。

图2 中国智能算力规模及预测，2019-2026



来源：IDC，2022

图3 中国通用算力规模及预测，2019-2026

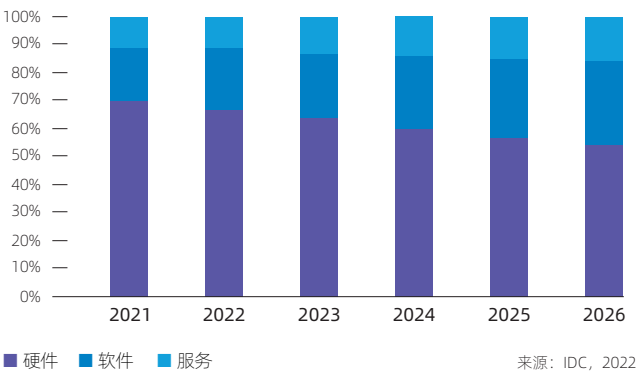


来源：IDC，2022

随着数字基础设施在政务服务、经济建设、民生保障、社会治理等方面的支持作用加大，以及伴随人工智能算力需求的高增长，中国政府鼓励推进数字公共服务的普惠化发展，建立健全助商惠民的数字基础设施服务体系，推进算力基建化发展。2022年2月，国家发展改革委会同中央网信办、工业和信息化部、国家能源局等部门，启动实施了“东数西算”工程。“东数西算”工程的启动以及智能计算中心的建设从国家层面实现有效的资源结构整合，助力产业结构调整，通过算力基础设施从点到网的升级，构建更为健全的基础设施结构。目前，国家在京津冀、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，协调区域平衡化发展，推进集约化、绿色节能、安全稳定的算力基础设施的建设。2022年发布的《企业技术创新能力提升行动方案（2022-2023年）》也提及将加速智能计算中心的发展，以期面向企业提供低成本算力服务，推进算力的基建化发展。

IDC调研发现，中国企业对人工智能算力基础设施平台的关注点依次为：丰富的应用场景配置、加速性能和计算能力、规模效应下的价格成本因素、训练的数据支持、人工智能配套政策吸引。**IDC认为，就现阶段而言，由于中国市场倾向于首先投资硬件，中国人工智能支出中硬件占比将保持最大，未来5年将一直保持65%左右的份额。**

图4 中国人工智能支出中硬件、软件、服务占比及趋势，2021-2026



来源：IDC，2022

企业在模型研发和落地过程中往往存在高投入、高风险等挑战，算法基建化可有效帮助企业实现破局。借助智能计算中心，企业可部署训练和推理系统，推进模型研发和创新，尤其有利于自然语言处理大模型、视觉大模型和多模态大模型等高算力消耗模型的构建。

除了大模型研发和创新，对于众多企业而言，他们还面临如何将大模型落地行业，解决现实复杂、琐碎场景中的应用问题。IDC调研显示，未来超过80%的组织表示会考虑购买预先训练好的人工智能模型，而不是自己进行训练。但是预先训练的模型在可用性和适应性、运行模型的基础设施，以及内部专业知识等方面还存在提升的空间，企业亟需行业的解决方案商的支持，缩小技术创新和落地应用之间的鸿沟。

人工智能算力及应用

2

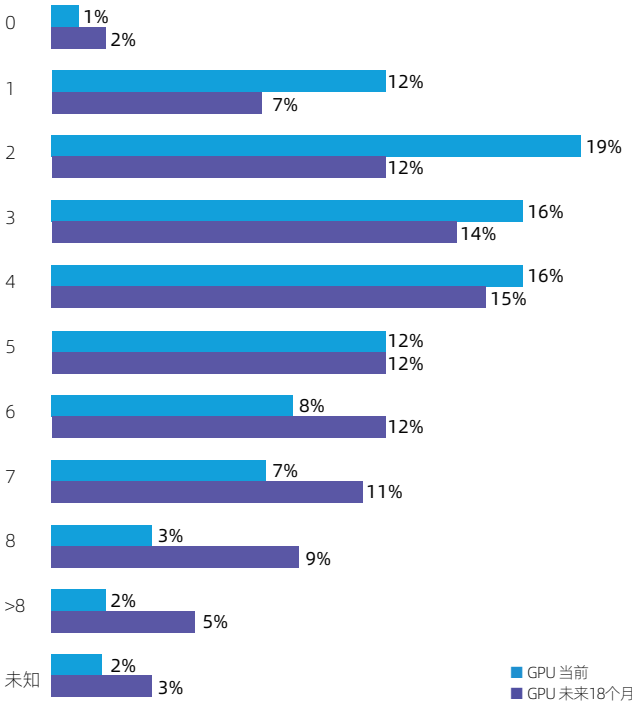
算力是实现AI产业化的核心力量，它的发展将对人工智能技术的进步和行业应用起到决定性的作用。随着人工智能向多场景化、规模化、融合化等高应用阶段方向发展，数据体量呈现出急剧增长态势，算法模型的参数量呈指数级增加，以加速计算为核心的算力中心规模将不断扩大。

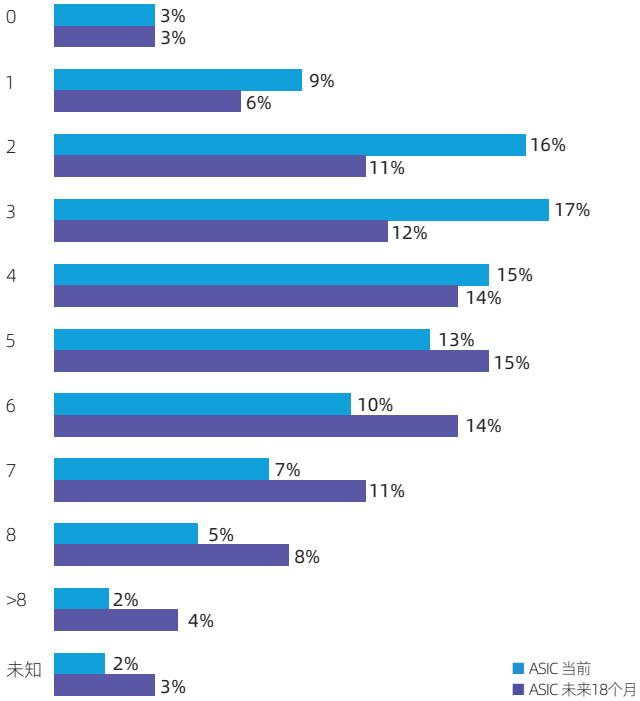
本报告将从算力基础架构层面，对人工智能芯片、服务器、计算架构、算法及应用等方面的发展近况逐一进行分析。

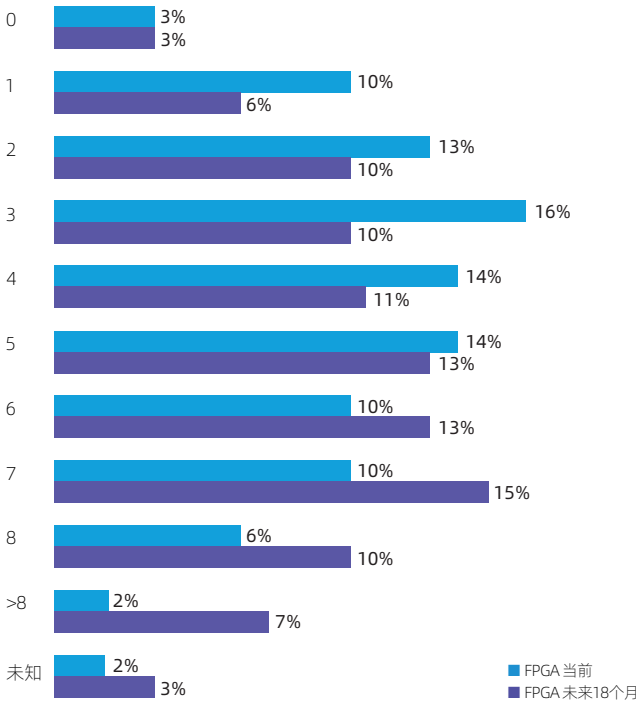
2.1 芯片：需求日益增长，发展空间广阔

全球人工智能技术发展逐渐成熟，数字化基础设施不断建设完善，人工智能产业技术不断提升，产业商业化应用加速落地，推动全球人工智能芯片市场高速增长，IDC预计，到2025年人工智能芯片市场规模将达726亿美元。IDC全球范围调研显示，人工智能芯片搭载率（attach rate）将持续增高，目前每台人工智能服务器上普遍多配置2个GPU，未来18个月，GPU、ASIC和FPGA的搭载率均会上升。

图5 全球人工智能服务器GPU、ASIC和FPGA芯片搭载率



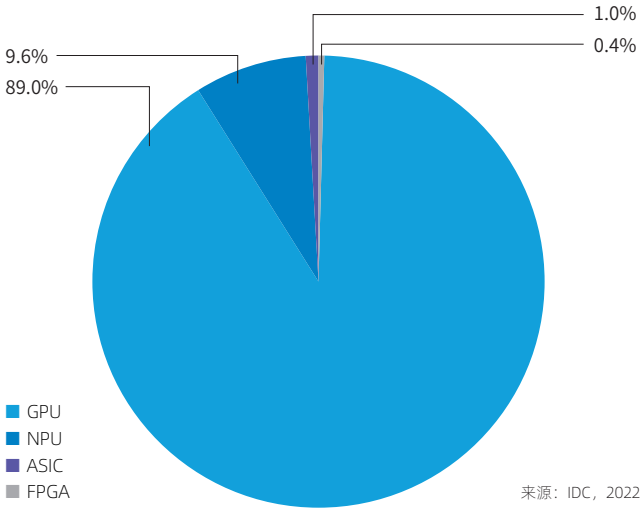




来源：IDC，2022

IDC研究发现，2021年中国仍以GPU为主实现数据中心计算加速，市场占有率近90%，GPU芯片多用于图形图像处理、复杂的数学计算等场景，可较好支持高度并行的工作负载，常用于数据中心的模型训练，也可以用于边缘侧和端侧的推理工作负载。ASIC，FPGA，NPU等非GPU芯片市场占有率超过10%，其中，NPU较以往具有明显增长，NPU芯片设计逻辑更为简单，常用于边侧和端侧的模型推理，并生成结果，在处理推理工作负载时，具有显著的能耗节约优势。

图6 中国人工智能芯片市场规模占比



在中国当前人工智能发展局势下，人工智能芯片产业发展体现出以下趋势：

- **人工智能芯片发展驱动力增强，未来发展前景广阔：**芯片产业不仅是信息产业的核心部分，同时也是国家信息安全的硬件保障，近年来中国高度重视芯片产业发展。随着政策支持以及资本推动等多重驱动，人工智能芯片不仅专利数量不断增加，产业链和应用场景持续完善扩充，人工智能芯片产业发展前景广阔。

- **不同类型人工智能芯片发展进度参差不齐：**在面向人工智能领域的芯片中，用于终端产品的应用层芯片发展较快，而用于云计算等领域的通用基础层芯片发展则较为滞后，有较大发展潜力。在近几年的热门领域如智慧城市建设、无人驾驶载具、智慧医疗系统构建、智能家居等应用中，ASIC、GPU、FPGA、NPU四大类芯片，受到了绝大多数中国人工智能产业链企业的青睐。

- **人工智能芯片低能耗为大势所趋：**低功耗人工智能芯片是时代之需，这对实现数据中心的总功耗降低具有重要价值，此外，低功耗人工智能芯片也是实现边缘智能的重要环节，能满足更多复杂、极端的边缘侧应用场景需求。但研发出能够满足性能需求，兼顾制造可行性、成本可控性、性能可靠性等要求的低功耗芯片仍颇具挑战。

- **产业分工成雏形，逐步完善生态建设：**企业面对多元化的算力和海量的数据，期待技术创新可以为企业带来活力。但由于芯片架构繁杂，开发工具匮乏，系统软件、应用开发平台配套少，应用难以迁移，资源不能复用和共享等问题使得生态复杂离散。面对这样的现状，芯片技术厂商（含设计、制造、应用等流程）应联合上下游厂商，建立技术生态系链条，加速人工智能场景的落地。

人工智能算力规模的快速增长将刺激更大的人工智能芯片需求。随着疫情、供应链及政治环境的影响，芯片领域的供需关系产生了很大变化，很多企业开始从“国际采购”转向“本地采购”，这对国产芯片厂商来说迎来了较好的发展机会，但整体来说，目前中国大部分基础芯片主要依靠国外厂商的供应，国产芯片厂商在全球产业链中不具有明显优势。本报告认为，中国芯片产业主要面临两方面的挑战：技术与国际先进水平相比相对落后，产业生态建设同样不够完善。中国将持续优化芯片产业发展环境，不断促进设计、封装等环节发展，对流片制造环节实现攻坚，加大对高通用性、高性能、高效率芯片的研发力度，以更好地支持如图像渲染、机器学习等重要人工智能应用场景。此外，中国还需构建健全完整的产业链，打通行业应用、芯片研发、系统开发、高校研究之间的壁垒，形成跨企业、跨领域、跨行业的合作，推进芯片行业全维度发展。

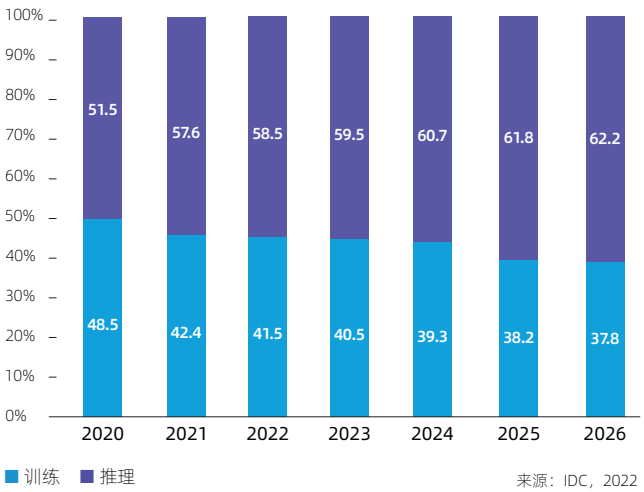
2.2 服务器：中国市场领跑全球，绿色节能引领未来

人工智能服务器仍是人工智能市场增长的主力军。IDC数据显示，2021年全球人工智能服务器市场的同比增速超过全球整体人工智能市场的增速，是整体人工智能市场增长的推动力。IDC发布的《全球人工智能市场半年度追踪报告》显示，2021年全球人工智能服务器市场规模达156.3亿美元，约合人民币1,045亿元，这是全球年度人工智能服务器市场首次突破千亿元人民币，同比2020年增速达39.1%。其中，浪潮信息、戴尔、HPE分别以20.9%、13.0%、9.2%的市占率位列前三，三家厂商总市场份额占比达43.1%。未来五年，人工智能服务器市场将继续高速增长，预计2026年全球人工智能服务器市场规模将达到347.1亿美元，五年复合增长率为17.3%。

在中国，人工智能应用的加速落地很大程度推动了中国人工智能服务器的高增长。根据IDC数据，2021年中国人工智能服务器市场规模达到59.2亿美元，与2020年相比增长68.2%，其中，浪潮信息、新华三、宁畅、安擎、华为等诸多中国厂商正加速推动人工智能基础设施产品的优化更新，探索赋能技术升级，为人工智能技术的用户带来价值。IDC调研显示，超过80%的中国企业将在未来一年持续增加人工智能服务器的投资规模，中国人工智能服务器市场将在未来五年保持稳定增长，预计到2026年，中国人工智能服务器市场规模将达到123.4亿美元。

- **从工作负载角度而言：**IDC认为企业将更多地使用人工智能服务器处理推理工作负载。伴随企业人工智能应用成熟度逐步递增，企业将把精力更多从人工智能训练转移到人工智能推理工作负载上，这意味着人工智能模型将逐步进入广泛投产模式，这将对企业的人工智能基础设施规划带来影响，企业需要更好地制定运营支出规划，提升服务器利用率。据IDC数据，2021年中国数据中心用于推理的服务器的市场份额占比已经过半，达到57.6%，预计到2026年，用于推理的工作负载将达到62.2%。

图7 中国人工智能服务器工作负载预测，2020-2026



来源：IDC，2022

- **从部署位置而言：**产业侧对于低时延人工智能服务应用需求递增。相较于科研、重型产业能够通过大模型、高密度人工智能计算满足需求的场景，便捷、低时延的人工智能应用场景愈发普遍。越来越多的数据将在边缘位置进行收集、分析等操作，并可被移动到数据中心以进行进一步价值挖掘。越来越多的企业将构建跨本地数据中心、云、边缘的全链路人工智能基础设施，形成一个包括数据收集、分析、汇总和存储等所有环节的人工智能战略。

- **绿色节能化发展：**人工智能服务器将朝着绿色节能的方向发展，实现低功耗、高效率的计算。“东数西算”工程在全国启动，通过建设国家算力枢纽，规划设立10个国家数据中心集群，朝着全国一体化大数据中心体系迈进了一步。此外，国家对节能减排也提出了更高的要求，在发布的《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》中提到，到2025年，国家枢纽节点的PUE要进一步降到1.25以下，对建设绿色、低碳的数据中心提出了新的要求。企业开始将重点转移到液冷技术的探索和应用，液冷技术不但有更高的散热效率，还能节约大量电能，目前已经有大量成功案例。不论是从政策角度，还是市场需求角度，抑或技术成熟角度，液冷都将成为数据中心的发展方向。

2.3 计算架构：以系统创新为基础，支持多元算力发展

深度学习是典型的多迭代计算类工作负载，庞大的计算力是催生业务价值的必要条件。在通用算力技术演进节奏放缓的大背景下，针对特定问题或特定领域来定义计算架构成为市场的普遍诉求，基于DSA（Domain-Specific Architectures）思想设计的人工智能芯片，在特定人工智能工作负载上表现出远超通用芯片的处理能力，大大推动了人工智能芯片的多元化发展，并为产业AI化的加速提供了重要的产业基础和更加丰富的选择。

然而，多元算力从“能用”到“好用”并且为企业创造业务价值，还有比较长的路要走，尤其是以百花齐放的AI算力芯片为核心，打造出一个通用性强、绿色高效、安全可靠的计算系统，对于推动人工智能技术普及应用至关重要。

一般来讲，从芯片到计算系统，需要完成体系结构、信号完整性、散热、可靠性等大量系统性设计工作，涉及到材料、热力学、电池技术、流体力学、化学等众多学科。由于人工智能基础架构往往是高密度集成的大算力系统，系统功耗、总线速率、电流密度等指标随业务需求持续攀升，给人工智能计算系统设计带来严峻挑战。业内从多个层面推动多元算力系统架构创新，充分发挥出多元化算力的体系创新优势，让算力好用、易用。

- **在系统层面**，由于人工智能芯片发展呈现多元化趋势，各厂商采用不同技术路线，产业面临硬件体系孤岛和生态割裂问题。加速人工智能技术产业发展，系统级产品创新是关键——在基础硬件、基础软件、核心应用、上层生态间建立起统一的技术路线及标准API接口，将加速器模块标准化，简化人工智能基础架构设计，缩短硬件开发和产业赋能周期。浪潮信息开放加速人工智能服务器NF5498，支持UBB v1.0 OAM基板，OAM兼容性高、扩展性好，支持多品牌异构加速芯片，已经在众多客户场景里面实现了落地，有效支撑国内外多元算力芯片发展。

- **在异构协议层面，**为了提高CPU与多元算力芯片间的数据传输效率，业内在互联技术方面展开了新的探索，近年涌现了一系列新兴的互连协议标准，包括QPI/UPI、CXL、GenZ、CCIX等，其中浪潮信息研发了支持CXL高速总线的智能加速器F26A，与传统的PCIe、DMA方式相比，CPU与加速器之间的平均数据访问延迟降低80%，同时可扩展2倍的内存容量。

- **在跨节点层面，**节点之间的网络通信所产生的RPC、协议处理、内存拷贝、压缩会占用30%左右的CPU资源，成为数据中心级的“通信税”。业内尝试通过智能网卡卸载计算密集型业务，将NVMe-oF、无损网络能力等功能转移到智能网卡上由专有硬件负责处理，能够提升通信性能并降低CPU占用率。通过智能数据处理单元和高速网络形成分布式互连交换，可实现CPU与各种加速芯片的算力协同以及内存池化、新型存储池化，节点间的数据访问延迟可低至亚微秒级别。

2.4 云服务：市场规模稳步提升，算力设施提供强力支撑

云计算的出现为企业提供更丰富的算力支持。通过aaS（as a Service）服务提供AI平台和AI服务，因其快速的产品迭代能力和丰富的场景化人工智能能力，越来越被用户接受。2021年全年，人工智能公有云服务市场规模达到44.1亿元人民币，占整体人工智能软件市场的13.4%。从年增长率来看，人工智能公有云服务市场的增长速度仍然远远超过人工智能软件整体市场的增长速度。而在未来2-3年内，IDC还观察到私有化部署仍将是整个人工智能市场的主流。

IDC调研显示，排名前三的人工智能云服务是： 搜索、人脸识别和推荐引擎，预计未来18个月，排名前三的人工智能云服务将为：自然语言处理、图像识别和视频识别。近几年，人脸与人体识别已经达到一定的市场规模，相比2020年，2021年人脸人体公有云服务市场规模实现80.1%的增长，应用场景的扩展，市场产品形态不断丰富，以及疫情防控等因素是重要驱动力。在图像视频领域，视频结构化、多模态人工智能等技术的创新促进了该领域市场增长。公有云厂商一方面通过视觉开放平台输出图像视频领域的人工智能能力，另一方面则专注于开发基于场景的解决方案。在自然语言处理方面，2021年NLP市场较2020年实现了126.9%的增长，在技术方面，得益于大型模型的推广；在市场方面，与应用场景的发展（如机器翻译、文档处理、智能写作等）息息相关。同时，智能语音公有云服务市场已经实现了高增长，2021年市场规模较2020年增长52.3%，目前已进入应用场景深化阶段。智能客服、客服质量控制、客服数据分析、智能营销等应用推动了对话式人工智能的市场增长，2021年较2020年增长109.6%。

2.5 算法模型：加速大模型行业落地，助力实体经济发展

大模型是在智算算力驱动下最为典型的重大创新。得益于模型泛化能力强、长尾数据的低依赖性以及下游模型使用效率的提升，大模型被认为具备了“通用智能”的雏形，并成为业内探索实现普惠人工智能的重要途径之一。大模型的技术基础是transformer架构、迁移学习和自监督学习，transformer架构应用于NLP领域并取得了突破性进展，其在视觉任务上也同样证明了有效性。从算力的视角看，语言类、视觉类模型容量和相应的算力需求都在快速扩大，大模型发展的背后是庞大的算力支撑。如果用“算力当量”（PetaFlops/s-day，PD），即每秒千万亿次的计算机完整运行一天消耗的算力总量，来对人工智能任务所需算力总量进行度量，AI+Science领域的AlphaFold2、自动驾驶系统、GPT-3等模型训练需要几百甚至几千PD的算力支持，如GPT-3训练需要3,640PD的算力。

2022年，大模型正在成为AIGC领域发展的算法引擎。在大模型的能力加持下，包括以文生图以及虚拟数字人等AIGC类应用将快速进入到商业化阶段，并为元宇宙内容生产带来巨大的变革。大模型正在让人工智能技术从五年前的“能听会看”，走到今天的“能思考、会创作”，未来有望实现“会推理、能决策”的重大进步。

大模型的发展同样给算力带来巨大的挑战。大模型训练的计算和存储资源开销之大，对加速计算系统和人工智能软件栈都有很高的要求，训练千亿、万亿模型动辄需要上千块加速卡，对大模型的推广和普惠带来了很大的挑战。同时，受限于边际递减效应，模型复杂度与精度的进一步提升将会需要更大比例的计算资源开销，对计算效率问题的顾虑会限制大模型参数规模的持续扩张。

尽管目前的大模型参数数量还没有达到人脑神经系统的突触规模，但市场对于大模型的认知趋于理性。业内逐渐认识到，大模型的发展更要注重绿色低碳、服务能力下沉以及商业模式的实践，为大模型在各行各业的规模落地铺平道路。

- **在学术界，**除了对预训练大模型进行性能优化，有大量研究集中在提升大模型的落地能力上，包括降低模型在预训练、适配下游任务和推理过程中的算力开销，通过模型压缩、剪枝、蒸馏等方法加快模型部署效率。

- **在工业界，**互联网企业推动自研大模型在电子商务、社交网络、搜索引擎、广告推荐等重点业务场景落地，并将内部大模型部署的成功经验固化为预训练平台，以标准化服务的形式将预训练模型的能力下沉到各行各业，通过集中式的数据和算力开发模式提供平台粘性，以期实现人工智能普惠化目标。加速硬件供应商提供分布式加速计算集群解决方案，通过并行算法与计算资源的合理匹配，提升加速计算集群整体利用率和大模型的训练效率，并优化推理加速库，以最少的计算资源达到最好的推理服务速度。

同时，业内对大模型的使用和选择也更加灵活多样，尤其在产业实际落地阶段，不再追求模型参数和算力的堆砌。业内通过知识蒸馏、模型裁剪、模型压缩等技术，基于通用大模型生成具备该行业或场景所需特定技能的专业模型，在保留通用大模型的知识、认知推理能力及泛化能力基础上，实现了针对该领域的技能专业化、模型轻载化和调用标准化。例如浪潮信息发布的四个技能模型——知识增强的对话模型、知识检索问答模型、中英文翻译模型、古文理解模型，在继承“源1.0”大模型通用的知识与能力基础上，面向特定领域的场景进行针对性的技能优化，模型精度和训练效率均处于业界领先：在十分之一参数量的情况下，即可在相同任务上复现98%的通用大模型效果，推理速度最高提升9倍。百度文心大模型覆盖各类AI应用场景和垂直行业，不仅包括NLP、CV和跨模态大模型，还包括生物计算和行业知识增强大模型，为医药、电力、金融和航天等各行各业智能化转型提供支持。

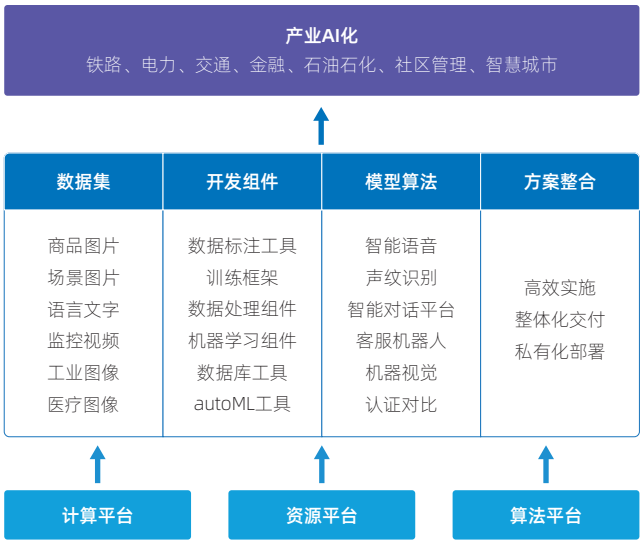
2.6 生态：推进产业化布局，发挥平台价值

技术创新的价值是提升效率，产业AI化的目标就是通过人工智能技术的应用来提升垂直行业研发、生产、运营等环节的效率，并产生更大范围的经济价值和社会价值。不过，人工智能目前还是新兴技术，技术供应商在实施产业AI化时仍面临诸多挑战，包括市场对人工智能技术的理解程度、供应商技术和商业能力是否成熟、如何在精细化落地的基础上实现快捷高效的部署，都会影响人工智能技术在产业内的实际落地效果。

通常来讲，推动某一类新兴技术应用走向成熟的基本路径和逻辑，是在产业发展初期，通过协同平台对多元市场主体的技术、产品、方案和服务等创新要素进行有机融合，并基于创新要素间的自由流动和非线性相互作用，激发出市场主体的创新动能，最终达到最优生产效率并产出符合目标需求的规模化应用产品。

对于人工智能产业，快速将创新要素转化为物质或知识资本，并形成规模效应和范围效应，关键在于协同创新平台的搭建。近年来，包括政府、企业、科研机构都在尝试构建人工智能协同创新平台，聚焦当前阶段产业AI化的落地应用需求，平台的存在可以更快实现人工智能生态伙伴的业务聚合、资源聚合和战略聚合，平台内的各方主体以人工智能算力输出、服务能力优化及人才培养等层面的要素供给，达成产业链上下游的通力合作形态，以生态聚合成就行业用户。

图8 人工智能产业生态协同平台架构



来源：IDC，2022

标准化是技术规模化应用的必要前提，但对于目前的人工智能技术及基础架构来说，定制化的工作量依然很大，主要集中在包括多元人工智能芯片适配、人工智能算力资源管理和调度、数据整合及加速、深度学习开发环境部署等各个方面。以人工智能芯片为例，市场上存在各种类型人工智能芯片，互联标准各不相同，用户在使用这些人工智能芯片系统时会遇到系统适配、芯片驱动、互联互通、功耗管理、安全传输、易用性等各类问题，给用户在部署多元人工智能芯片算力系统时带来巨大挑战。这些非标准的工作无法快速复用，限制了人工智能算力的使用效率，不利于人工智能在各行各业的推广和应用。

围绕人工智能算力产品市场呈现出的这些突出问题，AI算力和算法的基建化和标准化，是新时期人工智能产业发展和企业战略布局重点。近年来，包括公有云厂商和服务器厂商，都在推出具有标准服务能力的人工智能异构计算平台和算

力、算法一体化的新型基础设施，一方面布局建设智算中心，统筹算力的生产、聚合、调度和释放，为传统行业数字化转型、区域产业升级和基础科学研究等需求提供算力服务、数据服务和算法服务；另一方面从定制化DSA芯片、人工智能计算系统、高速互联网络、大吞吐低时延并行文件系统、人工智能增强容器调度等层面，输出多年技术沉淀和实践经验，为多样化人工智能场景提供软硬一体解决方案。

2.7 应用：场景化落地纵深发展，加速智算力向创新力转化

中国人工智能算力为人工智能的持续创新发展提供支撑。

对于宏观层面而言，人工智能算力为国家创造力的发展带来实质性推进：

- **创新环境：**如“十四五规划”所讲，人工智能已然成为“事关国家安全和发展全局的基础核心领域”，所以中国将持续瞄准前沿领域的发展，补足自身在人工智能基础理论研究和算法研究、芯片研发、原创性模型和框架的研发和迭代等方面存在的短板和劣势，加速人工智能单点技术的研究和创新，以政策支持、行业落地和企业推进为支撑点，加速相关产业的发展和人才培养。

- **创新科研：**作为创新的源动力，科学研究是人类发展和社会变革最主要的推动力量。随着人工智能技术的快速发展，人工智能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，极大提高了科学研究的效率并加速科学发展的进程，包括生命科学、数学、化学等多个领域。这其中，人工智能算力的重要性不言而喻。与行业应用不同的是，人工智能在科研领域所需要的数据精准度更高、模型更复杂，对于算力需求也更大。因此，人工智能算力、算法、数据和平台的结合能够为科研创新发挥更大的作用。

- **创新产业：**据工业和信息化部数据显示，目前，中国人工智能核心产业规模超过4,000亿元，企业数量超过3,000家，领军龙头企业覆盖无人机、语音识别、图像识别、智能机器人、智能汽车、可穿戴设备、虚拟现实等诸多领域，已经在智能芯片、开源框架等关键核心技术取得重要突破。根据美国斯坦福大学《2021年人工智能（AI）指数报告》中国在2021年提交了全球一半以上的人工智能专利申请，2021年全球约三分之一的人工智能期刊论文和人工智能引用是由中国研究人员所贡献的。

对于企业而言，人工智能算力可为企业带来切实的创新成效：

根据IDC针对企业应用人工智能现状调研发现，企业利用人工智能应用获得了显著收益，尤其是在研发速度和流程的创新，产品和服务的创新，以及决策制定的创新等维度：

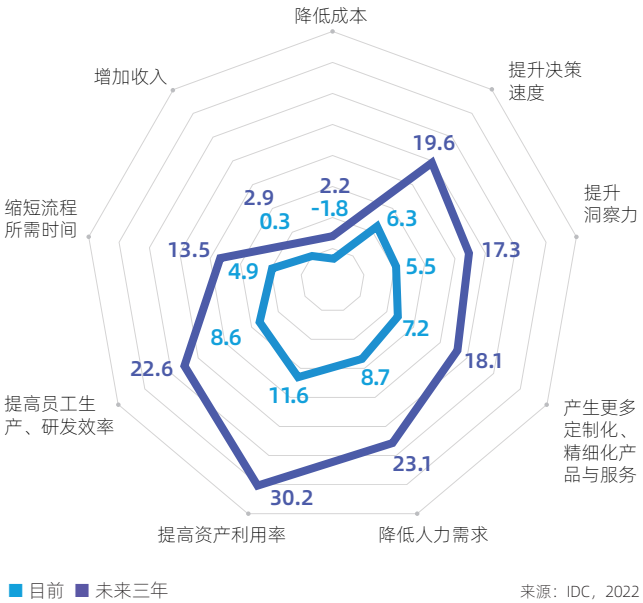
- **研发速度和流程：**根据 IDC 针对企业应用人工智能及算力现状调研发现，目前借助人工智能资源利用率平均提升12%， 员工生产和研发效率增加8%。

- **产品和服务：**人工智能为企业开发更多定制化和精细化产品与服务方面带来显著支撑，未来三年，有望在提高资产利用率，降低人力需求等方面更显著获益。

- **决策制定：**人工智能为企业带来更丰富、及时的信息，为企业决策者提供敏锐洞察，可显著提升决策的速度和质量，辅助企业处理复杂的不确定性，占领市场先机，创造价值。

根据IDC调研，尽管在过去一年，人工智能未在降低成本方面带来显著成效，但可以预计，伴随人工智能能力普惠化、规模化落地，未来企业有望实现降本增效创收的目标。

图9 人工智能目前及未来三年对企业产生的价值



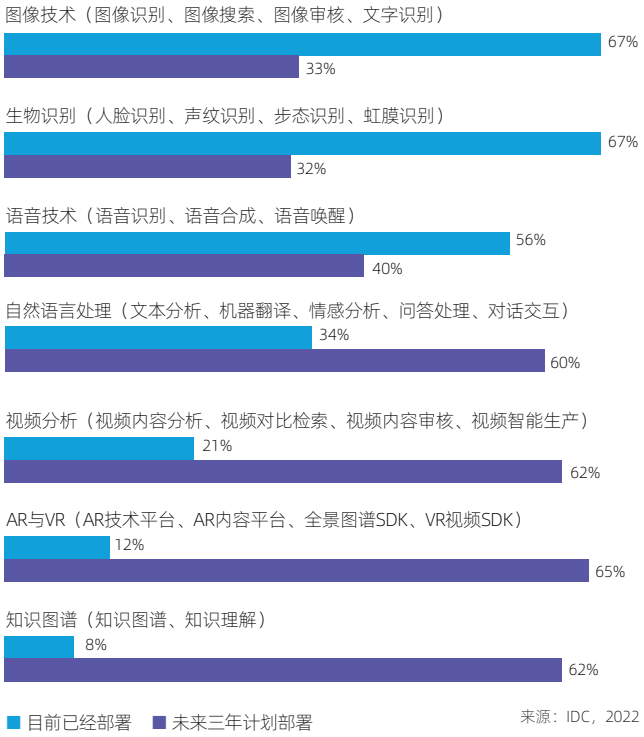
来源：IDC，2022

创新给企业带来了深远的影响，不管是超大规模企业还是中小企业，都在寻找适合自己的人工智能应用方式，并从中受

益。这也是全球企业在人工智能支出上持续快速增长的原因。IDC预测，2022年，全球企业将在人工智能解决方案上投资1,180亿美元。2021至2026年间，预计该支出将以26.5%的复合年增长率增长至3,010亿美元，这是同期全球IT总支出五年复合年增长率6.3%的四倍多。

从人工智能单点技术应用来看，根据 2022 年 IDC 针对企业对于人工智能技术的应用现状调研的结果来看，计算机视觉目前仍为最主要的应用技术类型，图像识别、生物识别、语音技术是目前较为广泛采用的技术，未来三年，AR与VR、视频分析、知识图谱和自然语言处理将成为主要发力点。

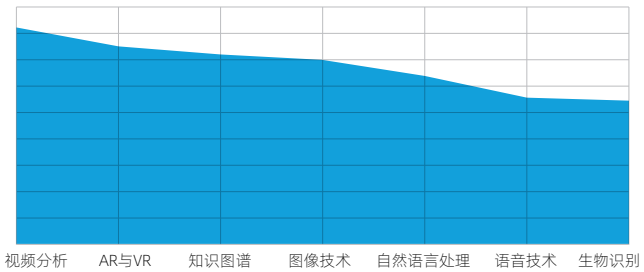
图10 企业已部署及未来三年计划部署的人工智能单点技术，2022



来源：IDC，2022

目前来看，视频分析、AR与VR以及知识图谱等需要基于强大的算力来满足诸如渲染、实时视频流分析、复杂计算等场景的需求，是目前企业主要的三个高算力消耗单点技术。

图11 单点技术场景对服务器资源的占用情况



来源：IDC，2022

从技术的行业应用而言，创新应用场景逐步增多。过去一年，中国人工智能应用保持快速发展的势头，行业应用场景相较去年也更加深入和细化。除了相对成熟的应用场景之外，物流、制造、能源、公共事业和农业等在人工智能的应用方面得到快速发展，创新应用场景逐步增多。

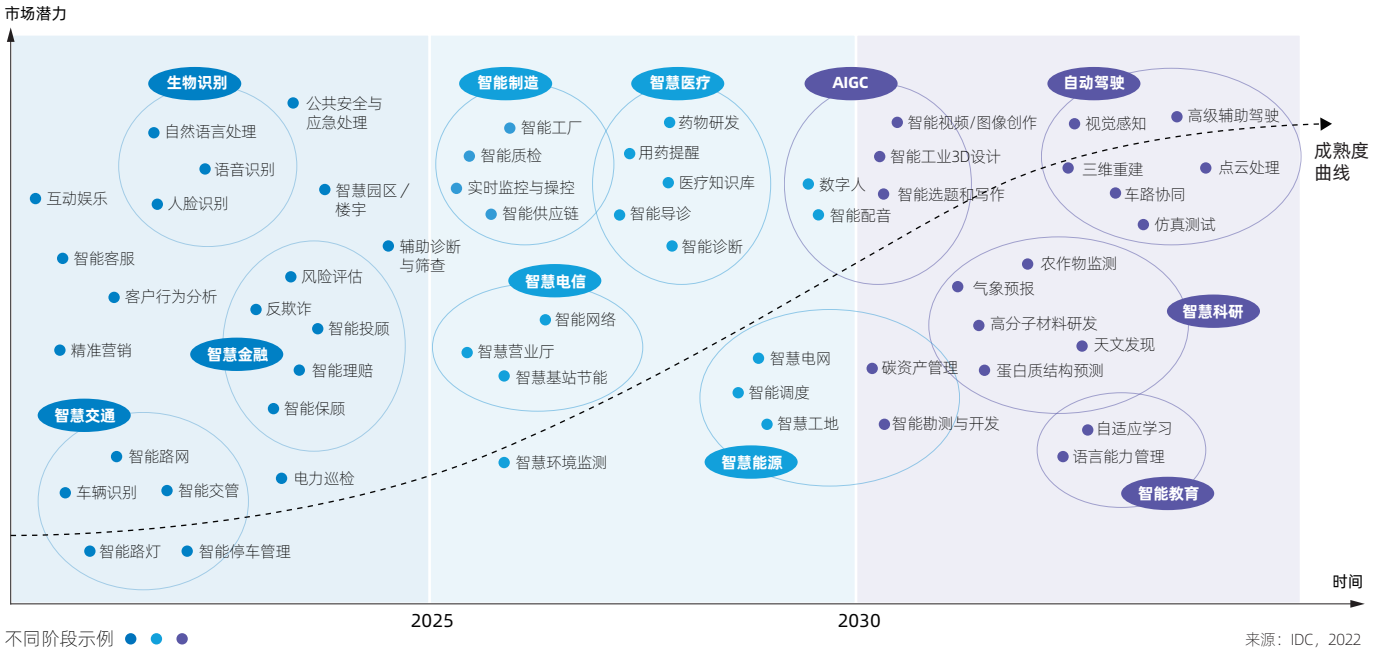
- **物流业**具有巨大的体量和较单一的运作模式，正在积极发展智慧物流。近年来，随着“造车热”的兴起，无人驾驶技术有了较大提升，目前技术可应用于短途或封闭环境下的物品配送。未来，随着无人驾驶技术的成熟和物联网带来的数据积累量的增加，人工智能将实现更广泛的应用，包括以移动机器人作为承载平台的无人仓储物流和实现产品生命周期全过程的高效协同的智慧供应链。

- **制造业**在过去几年对人工智能的投入增长相对缓慢，主要原因在于很多制造企业对投资回报率的把控严格，即使大部分制造企业认可人工智能所能带来的长期收益，但往往又妥协于企业的短期投资回报率。未来人工智能得以更多地渗透在制造企业的生产和设计流程中，制造企业中的人工智能的应用场景也会更加丰富，更加成熟，包括产品分拣、QC自动化、设备维护自动化、物流和供应链管理自动化等，IDC认为这些用例会促进越来越多的制造企业采用人工智能，制造业未来具有较大的人工智能发展潜力，更多的制造企业将利用人工智能提高自动化水平以及核心竞争力。

- **能源和公共事业**主要应用场景包括质检/巡检、流程运营自动化、预测性维护、供应链管理等，行业改革加速了智能化改造。以电力为例，在发电环节，与人工智能密切相关的智慧电厂的建设正在加速，对内通过智能化手段提升运营管理、内控效率，优化改造全生命周期；对外借助人工智能技术，对整个电力系统进行预测，匹配发电量，并对碳排放情况做估量，提升碳资产的管理效率。在电网层面的输电环节，虚拟机器人和无人机开始被广泛使用，针对所采集数据进行智能化诊断，利用人工智能应用不断提高故障诊断的准确性和反应速度。

- **农业**在过去几年对人工智能的投入增长较快，主要原因在于劳动力成本逐渐提高，农业种植逐渐规模化、集中化，且农业本身复杂度和危险系数较低，对于人工智能的要求并不高，目前人工智能已能胜任农业生产领域的大部分工作。未来，随着农业集中化的进一步发展和物联网积累数据量的增加，人工智能在农业中的应用将更加普及、场景也将更加丰富，包括农场种植采摘的无人化作业、养殖种植的气象土壤数据收集、作物的健康状况管理等，IDC认为这些用例会促进农业企业采用人工智能，因此，农业在未来具有较大的人工智能发展潜力。

图12 中国人工智能应用场景发展，2022



来源：IDC，2022

智能化场景在行业的落地呈现出更加深入、更加广泛的趋势：人工智能持续为提升用户体验做出贡献，诸如智能客服、智能推荐、精准营销等场景深入落地到各行各业；企业有意在数字人、虚拟NFT等数字化营销内容创作领域布局，以创造差异化的营销体验，升级品牌形象；此外，人工智能也在实现精准科学防疫，加强公共卫生安全体系建设中承担重要角色，在防疫信息汇总、病毒演变预测、疫苗药物研发、辅助诊断等维度实现广泛应用；人工智能正在加深对实体经济的支持，产生一批成熟应用的场景，包括但不限于人员设备管理、行为预测、供需销售预测等。另外，科学家们越来越多地利用人工智能技术和方法，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、深空深海等领域加速对前沿科学问题的探究。例如，在材料领域，科学家基于人工智能网络模型和大规模分子数据集，提升分子动力学模拟的极限，以快速、准确的方式预测新材料的特征，诸如预测高熵合金声子热导率随温度的变化关系、探究金属锂的自修复机理等；在数学领域，以往诸多定理的出现往往依靠求解者的直觉，而人工智能在数学中的应用可以加深数学研究者对于数学问题与现实应用中的关系，为提出新的数学定理并进行验证提供支持，以矩阵乘法为例，DeepMind利用强化学习的智能体AlphaTensor发现了超过2阶的高效的矩阵乘法，且比已知算法更快，推动新矩阵乘法算法的自动发现；在生物医药领域，科学家可利用人工智能高效预测蛋白质分子结构，提升预测速度和精度，对基础研究、生物制药和疾病诊疗具有重要意义，以某AI医药研发初创企业为例，它建立了药物发现平台，包括靶点发现和多组学数据分析、AI分子生成、设计和临床试验结果预测等功能，通过在并行计算机集群上开展AI辅助药物发现，仅用18个月和260万美元的投入便研发出特发性肺纤维化疾病治疗新靶点，而传统靶点药物研发一般需要4年以上时间。

未来五年，随着人机交互、机器学习、计算机视觉、语音识别技术达到更为成熟阶段，人工智能应用将呈现出如下发展趋势：从单点技术应用迈向多种人工智能能力融合、从事后分析迈向事前预判和主动执行、从计算智能和感知智能迈向认知智能和决策智能，以知识为主要生产工具的创作型工作（如文字、视频、图像和音频创作，软件开发，IP孵化等）将实现更大程度的智能化；行业企业也将持续创新，拓展数字孪生与人工智能技术的融合应用，推进在能源电力、制造、建筑等行业发展，构建虚拟工厂、数字孪生电网、数字孪生城市，加强数字与现实世界的连接，优化流程，实现全域管理，决策智能。

伴随技术进步对于人工智能在企业市场中的应用与落地带来促进作用，用于支撑应用的智算力已成为未来创新的核心保障。算力是数字经济时代的核心生产力，以人工智能为首的新兴技术应用在数字经济发展中起到了重要的作用，用于支撑人工智能应用的智算力决定了创新力的实现。不管是新型场景还是成熟场景，对算力都提出了极大的挑战，率先布局智算力的企业将在未来竞争中获得优势。

中国人工智能计算力发展评估

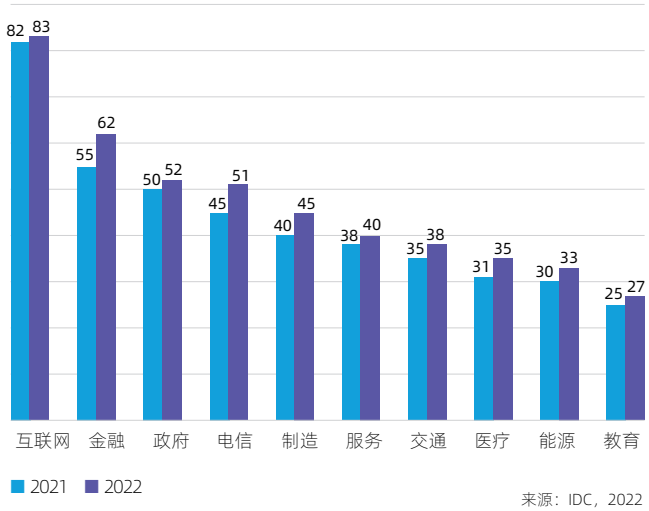
3

3.1 行业排名

总体来看，人工智能在各个行业的应用程度都呈现不断加深的趋势，应用场景也越来越广泛，人工智能已经成为了企业寻求业务增长点、提升用户体验、保持核心竞争力的重要途径。人工智能行业应用渗透度排名TOP5的行业依次为互联网、金融、政府、电信和制造。其中金融和电信行业人工智能应用增长速度较为明显，对人工智能基础架构的投入增长也较为突出。

2022年中国人工智能行业应用渗透度分布如下图所示：

图13 中国人工智能行业渗透度，2022 vs 2021



来源：IDC，2022

■ 互联网：人工智能技术研发和应用的排头兵

虽然在经历过去几年快速发展后，互联网在人工智能投入的增速有所放缓，但在中国，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。互联网企业普遍重视人工智能技术的研究和价值，并加速商业化落地。泛娱乐被公认为是互联网发展的重点领域，以IP价值挖掘为核心，电影、电视、动漫、游戏、音乐、网文等多领域逐渐打通，不仅行业的运作模式逐渐趋于成熟，对数字技术的接受和应用程度也处于领先地位。互联网企业在人工智能领域较早布局，随着互联网的深入发展和业务的多元化，在自然语言处理、知识图谱、用户理解、计算机视觉、语音语义、深度学习等技术领域都有较大的投资和积累，除了将人工智能应用到自身的产品服务和运营职能等方面，还将人工智能通过云服务平台对外提供人工智能服务。

在企业自身应用层面，互联网企业将人工智能应用在身份验证、自动化客服、营销互动、精准营销、舆情管理、内容审核等多个应用场景。例如，阿里巴巴已将各种人工智能技术广泛应用于诸多业务模块，以优化消费者体验，提升商业运营效率，包括基于深度学习和数据分析的个性化搜索结果和购物推荐、在搜索功能中采用语音识别和图像/视频分析技术以及智能客服，此外还将人工智能能力应用到人工智能音箱天猫精灵产品中。人工智能作为百度的战略核心之一，已被运行于百度的服务和应用中，包括语音助理平台DuerOS、自动驾驶平台Apollo、百度云人工智能解决方案和云服务、百度搜索和百度Feed等；通过百度人工智能开发平台，百度向第三方开发者开放了百度人工智能功能并提供了百度云上的工具包。除了超大规模互联网企业之外，其他互联网企业也不断加速人工智能应用，例如视频类互联网公司将人工智能技术广泛应用于大数据推荐等场景，并加速机器学习平台的构建，基于基础架构和算法投入，以及人员成本的综合考量，在诸多场景通过机器学习平台，可极大提升效率并节约成本。

■ 金融：持续优化用户服务和风险把控能力

人工智能在金融行业的应用增长迅速，位于第二位。人工智能和金融行业的紧密结合，为客户带来更好的体验感，为客户提供更大程度的便利。智能客服、实体机器人、智慧网点、云上网点等是目前人工智能在金融行业的应用典型。

此外，欺诈一直是长期困扰金融业的问题。人工智能的出现让信用卡公司可以在欺诈检测工作流程中实现预测性分析，人工智能可以驱动一系列分析工具，通过学习和监控用户的行为模式来收集和分析数据，以此判断是否存在欺诈风险，并显著减少误报情况。

在贷款信用评级方面，目前许多金融科技公司和金融机构已经开始向客户在线发放贷款。虽然这种方法为客户提供了便利，但在批准任何交易之前，检查客户的财务背景及风险状况至关重要。人工智能可以在优化在线贷款方面发挥关键作用。金融科技公司可以利用人工智能，根据机器学习算法提供的预设的数据分析和相关模式来检查客户的金融背景，根据客户的风险状况准确地分析客户。人工智能可以帮助实现在线贷款审批流程的自动化，还可以帮助金融科技公司批准预先确定的贷款金额。除此之外，金融科技应用程序的研发人员可以在人工智能和机器学习技术的基础上，加入更多的功能，如EMI计算器和贷款资格的自我评估。通过简化流程和提高服务质量，人工智能和金融科技的融合将给企业带来新增长。

最佳实践

招商银行：推进云原生时代的智能化发展

1987年，招商银行成立于中国改革开放的最前沿——深圳蛇口，是中国境内第一家完全由企业法人持股的股份制商业银行，也是国家从体制外推动改革的第一家试点银行，现已发展成为沪港两地上市，拥有商业银行、金融租赁、基金管理、人寿保险、境外投行等金融牌照的银行集团。招商银行一直走在数字化转型的前沿，高度重视数字化转型工作，并不断对外输出，为社会赋能。

为积极应对行业发展过程中面临的挑战，招商银行加快推进金融科技转型，推出了体系化的线上服务，一举覆盖客户多种经常性业务需求；重视发挥大数据和人工智能等新兴ICT技术的价值，以科技敏捷带动业务敏捷，深度融合科技与业务，快速迭代、持续交付产品和服务，优化客户体验。目前人工智能已经在诸如智能客服、智能风控等维度发挥重要作用。浪潮AI服务器作为底层基础设施，为提升AI算法复杂度和精度提供有力支持，建立高性能可扩展的事中实时风控架构，为智能化场景的落地奠定良好的基础。

招商银行在加速自身数智化转型的过程中非常重视利用云原生的技术优势。目前，招商银行的数据中心基本已经实现云化改造。云化的过程中，人工智能技术的应用对架构提出诸多新的挑战。传统业务下，一般应用在拆分为数个微服务后，服务间的调度、通讯时，参数和通讯量并不大。而在智能时代，数据呈现出海量增长态势，诸如图片或者视频等非结构化数据增长，大模型、大参数也对架构提出了更高的要求。招商银行在构建人工智能架构时重视不同集群的聚合性需求，探索出既不影响拆分层级的逻辑、又能保持弹性伸缩的架构方案，搭建出异构兼容，支持灵活调度算力资源，更好构建云原生时代下的智能基础设施。

伴随人工智能的逐步深入落地，招商银行认为实现算法模型规模和实际业务价值之间的平衡十分必要。在大模型训练后，企业需要考虑应用上线时的规模、业务价值创造和成本、资源消耗的平衡。根据实际模型的具体参数，进行一系列的压缩。随着规模下降，性能也会逐渐下降，但在此过程中，模型的真实业务价值和成本消耗也会发生变化，招商银行持续探索不同场景、不同算法模型、不同业务规模下的收益与投入平衡，与此同时，希望以最优的资源消耗方式获得更具效益、更加绿色的算力。

最佳实践

某国有大行：夯实基础能力，持续推进智能化发展

2022年1月，中国人民银行印发《金融科技发展规划(2022-2025年)》，其中提出要高质量推进金融数字化转型。人工智能业务创新发展需要有强大的算力基础设施支撑。近年来人工智能算法规模高速增长，传统人工智能集群1-2周的训练周期已经难以满足人工智能业务发展诉求。但仅仅提升服务器GPU性能，并不能带来算力水平的显著增加，单卡训练的性能瓶颈、算力资源的优化调度水平也亟待突破。

积极探索人工智能大规模并行训练创新实践

某国有大型商业银行为满足大规模智能化应用需求，进一步提升智能化算力水平，与浪潮信息合作打造“中高算力GPU并行运算集群”，构建领先的AI计算系统与智能业务生产创新平台，助力金融新业务新场景创新。

- 算力基础设施方面：引入高能效GPU算力，采用单节点8-16张GPU卡的中高密度算力节点。
- 网络基础设施方面：采用了100G高性能以太网络技术，较传统以太网络提升20%以上GPU集群训练性能。相较业界典型InfiniBand组网方案，具备更好的延展性，满足了人工智能集群训练大规模部署需求。
- 人工智能调度平台方面：依托浪潮AIStation智能业务生产创新平台，构建具备高性能、高可靠、可扩展的大规模GPU算力资源统一管理和人工智能作业调度平台，可实现对任意数量GPU资源组合的自动化调度，精确匹配不同规模人工智能分布式训练作业对资源的需求，提升集群算力的整体利用率。同时采用故障容错和断点续训技术，提升异常情况下人工智能训练的可持续性。

加速赋能金融业务数字化转型

夯实的基础设施建设为人工智能算力模型训练和智慧应用规模化建设打下了坚实的技术基础，带来的转变包含：

- 提升了性能和算力：中高算力GPU并行运算集群场景下采用高密度算力服务器及基于RoCE的高性能网络技术，相比传统算力提升8倍，相比传统以太网性能提升20%以上，相比其他高性能网络互连技术标

准具备更好的兼容性，能够更好地支撑中高算力GPU并行运算集群跨节点高速转发。

- 实现基础资源降本增效：相同模型训练，中高算力GPU集群不仅可以有效降低总能耗，同时可以减少机柜占用，提高集群算力密度，实现基础资源降本增效，为银行业务创新提供了有力的基础技术支撑。

此外，依托国内大规模客户群及产生的应用场景、数据和模型等，中高算力GPU并行运算集群场景下采用高密度算力服务器及基于RoCE的高性能网络技术有利于其研究金融高性能人工智能算力平台的关键技术，规模化建设不同型号的中高算力GPU服务器异构集群，并形成银行GPU算力共享资源池，寻找更契合银行发展的人工智能业务模型，进一步提升计算效率，支撑银行风控、营销、投顾等智慧银行场景。

通过开展中高算力GPU并行运算集群创新实践，该行模型训练效率显著提升。在金融凭证识别场景超大规模图片训练周期由数周缩短为1天，进一步缩减更新迭代时间，保持智能运营领域的技术领先。该行凭借企业级人工智能平台的先进技术能力及规模化应用情况，顺利完成数据处理、模型构建、模型部署、支撑与服务等评测任务，首批通过人工智能平台全能力域领先级评测，体现出该行在人工智能领域的领先优势和标杆示范作用。

■ 电信：优化网络构建，提升用户服务体验

人工智能技术已经成为电信行业不可缺少的部分，运营商凭借庞大的用户基数获取海量的数据，来源丰富、覆盖行业广泛、真实性高，形成了高价值的训练数据集，为人工智能在电信行业的未来发展奠定了基础。电信行业人工智能应用主要体现在两个方面：一方面，电信网络的构建及优化智能需要多项人工智能技术的融合，包括GPU加速、深度学习和分析技术等；另一方面，下一代智慧网络的打造，对云化网络的智能编排、调度、运营等也需要人工智能技术的支撑。另外，对运营商自身而言，越来越多的电信运营商着手于智慧营业厅的建设，捕捉消费者数据，比如停留时间、行为习惯等，然后利用人工智能技术对获取的信息进行分析并及时作出反应，如增加柜台客服人员或引导客户使用自助服务系统。电信行业的人工智能应用不仅优化和维护了基础设施、给客户创造了更好的体验感，还能提高业务营收、降低成本以及提高效率。除了自身应用之外，用于云服务的投入在电信运营商所占比重也逐渐增加，并利用多样化架构实现更加优化的运行，运营商人工智能相关投入有望在未来几年保持高速增长。

■ 医疗：赋能诊断治疗，加速科研探索

医疗行业虽然在人工智能应用层面起步较晚，但在最近一年有了显著提升。人工智能已经被中国的医疗机构和生命科学组织广泛接受，但因相关标准和规范还不完善，只有少量医疗人员参与人工智能开发和应用。在未来五年，随着中国政府对人工智能开发和应用相关法规的完善，人工智能在医疗行业的应用将会快速拓展。

在医疗数字化转型的过程中，传统医疗向互联网医疗模式的转型趋势愈加明显，而人工智能、大数据等新技术的发展则使得疾病诊断和治疗的模式发生转变，从单点开始逐步扩展到各个领域，带动了医疗信息化的全面升级。当前，医疗人工智能系统主要采用的技术包括计算机视觉、自然语言处理、机器学习等，目前应用场景主要分为三个方向：

- **文字方向：**医院沉淀了大量电子病历，不管是电子健康档案还是电子病历，都是以文字方式积累。人工智能技术能够帮助医院自动识别文字含义及上下文关系，建立对应的医疗知识图谱，用于辅助诊断、用药提示、科研挖掘等。

- **图像方向：**通过图像识别方式辅助医师检查，准确率已经达到较高程度。在抗击新冠疫情中，人工智能在疾病救治和疫情防控中发挥了巨大作用。例如，对患者的肺部放射影像诊断需要医生检查大量的放射影像，耗费大量的精力和时间，医学影像人工智能辅助诊断系统的应用极大提高了诊断的效率。

- **生物方向：**基因数据、基因组的数据非常庞大，大量医药企业正通过临床经验结合标志属性去挖掘发现更多的肿瘤标注，加速新药研发过程。相对文字和图像方向，生物方向类人工智能应用场景还处在相对初期阶段，在政策支持和市场需求的推动下，未来具有极大的发展潜力。

■ 制造：加速应用场景落地，赋能企业降本增效

人工智能是制造业迈向工业 4.0 和工业互联网时代的重要新兴技术能力。制造业对于人工智能技术的使用正在稳步上升。在制造业中人工智能不断丰富和迭代自身的分析和决策能力，以适应不断变化的工业环境，帮助企业在产生大量结构化和非结构化数据的复杂生产环境中更为快速、准确地梳理参数之间的相关性，提高生产效率，优化设备产品性能，具有自感知、自学习、自执行、自决策、自适应等特征。制造

业中的人工智能的本质是实现复杂工业技术、经验、知识的模型化和在线化，从而实现各类创新的工业智能应用。制造业在人工智能的主要应用场景包括：交互界面智能化、质量管理及推荐系统、维修及生产检测自动化、供应链管理自动化、产品分拣等。**IDC预计，到2023年年底，中国50%的制造业供应链环节将采用人工智能，从而可以提高15%的效率。**这将使企业能够更好地预测市场变化、消费趋势和习惯的变化，甚至是气候变化，进而将预测结果与库存管理相联系，帮助企业努力使库存水平贴近市场需求，促进销售，同时降低成本，把控风险。

此外，诸如媒体和娱乐、游戏、建筑等行业也在加速元宇宙技术的落地和应用，基于人工智能、物联网、智能边缘等技术，满足市场对于多元化、定制化、共情化的体验，改善运营流程，加速学习、分享、创造，产生更大的经济和社会价值。实现元宇宙构想以及物理与数字世界间的互联，需要创建更多的数字资产/数字人，这对计算性能与计算资源提出新的要求。目前元宇宙基础设施的搭建已经开始起步，通过构建能够支持应用落地的人工智能算力基础设施，提升基础平台的支撑力度，为将来满足企业和用户在虚拟环境中的应用需求夯实基础。

最佳实践

青田元宇宙智算中心：数实共生，探索创新

青田元宇宙智算中心坐落于浙江省东南部的青田县，是国内首个元宇宙智算中心，同时也是首个算力、算法、开发平台一体化的新型元宇宙基础设施，通过协同创建、高精仿真、实时渲染、智能交互四大作业环节，将为多产业的元宇宙场景提供算力的技术支撑，在实体数字化方面逐步突破多维度和多场景的屏障。

青田元宇宙智算中心以建立国家级产业标杆为目标，汇集1000+生态开发者资源，引导十万级AI、元宇宙产业精英和科技人才汇集，共同实现元宇宙与边缘计算战略落地，同时开展5G应用创新研究，在娱乐、工业等领域打造相关行业标准。

青田元宇宙智算中心的启动，将为包括青田县乃至浙江、长三角地区等多个地区的元宇宙产业提供技术支持，同时，青田元宇宙智算中心也会驱动地区经济发展，为产业数字化融合提供新的动力。

围绕新需求，构建新能力

元宇宙概念强调自制、共制、共享，在高度拟真的数字世界里实现海量用户的实时交互，这些目标的实现依赖协同创建、高精仿真、实时渲染、智能交互等多个重要环节。每个环节都需要巨量的人工智能算力作为支撑，这对人工智能算力基础设施提出更高的要求，除构建高性能、低延迟、易扩展的硬件平台外，还需要有端到端、生态丰富的软件栈的支持。

- 以协同创建为例，大规模、高复杂的数字孪生空间、数字人和其他实体角色的建模需要众多设计师协同创作完成，较好的底层平台虚拟化和云端协同能力可很大程度助力效率的提升。

- 在产业维度，更高效的建模技术依赖TB级的数据运算、精简3D模型所需数据量，以及物理维度算法的设计。为了更好地为元宇宙世界输入模型，可以针对垂直应用场景打造更优化的建模技术，从而节约虚拟建模的时间。同时，进阶的算法可以精简3D模型所需的数据量，从而提升软件处理的效率。通过运用基于现实物体物理材质建模的算法，以及元宇宙数字协同综合产业平台可以为产业提供高效的3D工具，实现开发者生态、软件生态以及垂直应用场景的正循环发展，拉动底层算力的使用效率。

青田元宇宙智算中心作为浙江省青田县人民政府与浪潮信息、谷梵科技三方共建项目，建成后每秒算力性能将超过10亿亿次。

- 硬件方面，青田元宇宙智算中心采用浪潮领先的异构加速服务器的旗舰系统，具有强大的RDMA通信和数据存储能力，可提供强大的渲染和AI计算能力。

- 软件方面，青田元宇宙中心，依托谷梵在AI数据中心和“行业+AI”创新赋能平台的架构设计、流程设计、功能组件开发和接口API开发上的能力，构建AI项目任务发布市场，实现车路协同、数字人、工业制造的供需对接，同时建立对3D设计协同、增强数字孪生和世界仿真、3D工具低代码开发等核心应用的支持，让设计师实现更大程度的自由创作。将元宇宙协同平台的强大的功能与浪潮元宇宙服务器的性能优势有机融合，联合打造强大的软硬件生态。

- 与此同时，发挥浪潮AIStation资源平台、算法平台和应用优化服务，以及“源”大模型的能力，为全行业全场景的元宇宙应用提供能力支撑，为用户打造高效的元宇宙协同创新体验。

为当地经济发展带来强劲动力

未来，青田元宇宙智算中心将：

- 成为成熟的数字资产交易平台，通过合作伙伴帮助搭建数字孪生场景，可以为如物流、服装业等生产

的数字孪生资产提供超高价值的参考借鉴用途，帮助这些产业快速降低同类型场景的搭建时间、搭建成本以及升级完善，最终推动行业类数字孪生场景发展，形成下一时代的跨行业和场景的工业元宇宙雏形。

- 将视觉产品代入市场，为视频内容创作（如短视频）提供更高的技术支持。通过打造元宇宙数字协同综合产业平台，可以将影视级产品所需的硬件平台共享给原本无法承担成本的创作者，可以大幅提升创作者的制作水平，帮助其推出影视级视觉产品的同时又符合市场高速迭代的需求。

青田元宇宙智算中心的建设，将为元宇宙初级阶段的落地提供普惠、安全的算力支持，创建跨越地理距离的虚拟世界，让艺术家、设计师、工程师和科学家实现实时3D仿真、设计协作，在此过程中探索出新的发展模式，促进青田人创新创业，同时在对应硬件的支持下，促使元宇宙从软件平台转变为工业集群，提供更优质专业的公共服务。

3.2 地域排名

本报告针对不同城市在人工智能投资规模（包括人工智能算力投资规模，人工智能其他投资规模，未来投资计划）、人工智能相关政策支持力度（包括人工智能相关政策扶持力度、政策落地情况和实施进展）、人工智能技术成熟度（包括人工智能技术应用成熟度、第三平台技术应用成熟度、数据平台成熟度），以及劳动供给（包括人工智能相关技术人员数量和水平、AI企业人数/企业数量、未来人才储备）等维度的情况，并基于持续研究和最新用户调研，进行综合评估。在2022年中国人工智能城市排行榜中，北京位居首位，杭州位居第二，天津进入前十名。除了TOP10城市之外，多个城市在自身产业优势及各种因素推动下，人工智能应用取得了较大进展，例如合肥、长沙和武汉等，中国人工智能城市发展正遍地开花，未来将会出现越来越多结合城市特点的人工智能示范区，为产业发展树立标杆。

图14 中国AI计算力发展评估——城市排行，2022

来源：IDC，2022

■ 北京

北京在人工智能领域的政策、人才、技术、企业等方面的优势有助于自身在人工智能领域取得领先地位，并将继续吸引新的企业和投资资金进入。北京作为我国首个国家新一代人工智能创新发展试验区，人工智能产业规模快速增长，据悉，截止2022年11月，北京已经聚集了约1500家人工智能领域的相关企业，在17个人工智能相关领域领跑全国，突破多项核心技术。《加快新型基础设施建设行动方案（2020-2022年）》，《北京市促进数字人产业创新发展行动计划（2022-2025年）》，《关于支持中关村科学城智能网联汽车产业创新引领发展的十五条措施》等多项针对人工智能发展的政策发布，推动人工智能技术在前沿领域的探索，在重点领域的突破。除了政策的支持以外，北京还拥有优质学术资源和人才资源的加持，与人工智能相关的人才占全国总量的一半以上；北京大学、清华大学、北京航空航天大学、中科院自动化所、中科院计算所等全国过半数人工智能主要研究单位都聚集在北京。

■ 杭州

2022年初浙江省政府发布了《建设杭州国家人工智能创新应用先导区行动计划（2022-2024年）》，指出到2024年，全市人工智能应用水平全国领先、国际先进；推动科研院所与龙头企业双擎联动创新，在人工智能基础理论与核心技术攻关上取得重要进展，形成10项以上人工智能重大科技成果，获得1000项以上核心发明专利；打造3~4个千亿级人工智能产业集群，产业营业收入年均增长15%以上，产业综合竞争力位居全国前列。在知识产权方面，人工智能学科建设在杭州历史悠久，早在1987年便设立了人工智能研究所，是国内最早的人工智能研究所之一。如今，在阿里巴巴集团和浙江大学、杭州电子科技大学、浙江工业大学等高校的加持下，杭州人工智能产业发展迅猛，人工智能专利授权数已超过千件，位居国内人工智能第一梯队。

■ 深圳

作为全国最年轻的一线城市，深圳位于珠三角、毗邻港澳，地理位置十分优越。此外，深圳是我国改革开放的重要窗口城市、是我国首个国家创新型城市。在国家政策的支持下，深圳吸引了众多外资和企业前来投资，落地了大量发达国家的高端产业；同时虽然本地高校资源不足，但每年吸引了大量人才前来就业，为深圳的发展提供了人才保障。目前深圳人工智能企业已超过一千四百家，更有华为、腾讯、平安等一众超大型公司加持，在智能语音技术、计算机视觉、自然语言处理等方面具有显著优势。

■ 上海

上海地处长三角城市群，是中国的经济中心，拥有丰厚的历史底蕴、开放的市场环境和完备的产业体系，在智能制造和智能交通等领域构建了丰富的应用场景。2021年，上海发布《上海市人工智能产业发展“十四五”规划》（简称《规划》），指出要深化人工智能在城市数字化转型中的重要驱动和赋能作用，加快建设更具国际影响力的人工智能‘“上海高地”，预计到2025年，上海将形成10大类100个人工智能深度应用案例，培育500家智能化示范企业。同时，上海市作为国内的“教育高地”，拥有复旦大学、上海交通大学等一众一流高校资源，为人工智能的发展提供了人才保障，目前上海市人工智能专利授权数仍处于全国领先地位，根据《规划》，到2025年上海人工智能人才规模将达到30万人。

■ 广州

广州作为我国的一线城市，2021年发布了《广州市人工智能产业链高质量发展三年行动计划（2021-2023年）》，指出构建广州市人工智能产业“链长制”，通过开展“十百千”战略发展计划，建设 10 个人工智能产业园，开展 100 个人工智能典型场景应用示范，培育 1000 家左右人工智能企业，创新协同良好的产业链体系。实施“2+4+N”产业培育工程，遴选先进制造、车辆交通、健康医疗、城市治理四条人工智能优势赛道，重点培育龙头企业和高成长性企业，赋能人工智能产业的发展。

■ 成都

作为一座历史悠久而独特、文化积淀深厚的城市，成都不仅物产丰富、农业发达，更是中国重要的电子信息产业基地。成都有国家级科研机构30家，国家级研发平台67个，高校65所。据2019年统计，世界500强企业落户成都的已超过300家。从2021年开始，成都开始建设国家新一代人工智能创新发展试验区，注重把握人工智能发展战略机遇。

■ 苏州

2021年，苏州发布了《苏州市促进新一代人工智能产业发展的若干措施》，开展具有行业引领性的人工智能“头雁”企业遴选，鼓励各地区加大对“头雁”企业及“头雁”培育企业的政策奖励。同时支持人工智能企业总部从外地迁入苏州，国内外人工智能知名企业在苏州设立子公司。鼓励企业、科研院所主动参与承接国家、省新一代人工智能产业创新重大项目。推动建立重点支持领域创新项目库，对技术水平先进、商业模式领先的项目给予项目扶持。

■ 南京

目前南京在扩大人工智能产业规模、培育优质企业、改善城市技术布局，以及提升应用平台支撑等多个方面正在实现跳跃式发展，着重于人工智能产业的基础层、技术层、应用层三大层级，汇集了人工智能企业近300家，其核心产业规模超过60亿元，同时，带动相关产业规模近800亿元。南京积极开展人工智能生态圈和产业链建设，发挥在图像识别和智能传感等领域的优势，基于智算中心的集群效应，促进诸如车联网、智能制造、自动驾驶、智慧医疗等领域的发展。

■ 天津

近年来，天津市紧抓人工智能产业发展机遇，通过举行三届世界智能大会，推动一大批人工智能项目落地。同时在政策方面，2020年天津市政府推出《天津市建设国家新一代人工智能创新发展试验区行动计划》，指出要加速重大应用场景落地，提升综合支撑力；加快培育人工智能产业，提升产业聚集力；加大人才引进培养力度，提升创新创业活力；力争到2024年，天津人工智能试验区建设取得显著阶段性成效，成为引领全市人工智能产业发展的核心载体。目前天津的人工智能产业链主要集中在上游的芯片行业以及下游的人工智能应用场景。全市累计支持八批2998个项目，带动投资超过1200亿元。建成200个智能工厂和数字化车间；打造了涵盖智能制造装备及产品、工业软件及控制系统、智能制造专业服务等领域的产业体系，成为支撑制造业立市的“新动能”。未来天津还将通过典型示范应用，吸引聚集上下游优势企业，形成国产基础软硬件产品群，打造完整产业链条。

■ 济南

济南是中国软件名城，软件与信息技术服务业规模总量占全省比重超过百分之五十；以服务器、超级计算和量子信息为代表的算力产业在国内外保持领先，浪潮信息人工智能服务器连续五年保持国内市场占有率第一。丰富的应用场景是济南人工智能发展的重要力量，在能源、医疗、智能制造等领域均有众多企业布局。知识产权方面，济南人工智能软件著作权共4500余件，平均每家企业拥有16件软件著作权；人工智能企业拥有发明专利1500余件，平均每家企业拥有5件发明专利；平台载体不断壮大，全市拥有人工智能产业创新平台53个，华为、百度的创新中心落地济南，加速构建人工智能产业发展新生态。

表1 近五年TOP10城市排名变化

排名	2022	2021	2020	2019	2018
1	北京	北京	北京	北京	杭州
2	杭州	杭州	深圳	杭州	北京
3	深圳	深圳	杭州	深圳	深圳
4	上海	南京	上海	上海	上海
5	广州	上海	重庆	广州	合肥
6	成都	苏州	广州	合肥	成都
7	苏州	广州	合肥	苏州	重庆
8	南京	济南	苏州	重庆	武汉
9	天津	成都	西安	南京	广州
10	济南	合肥	南京	西安	贵阳

来源：IDC，2022

除以上城市外，还有些城市深耕特定的人工智能应用并取得了明显成果，成为城市的智能化新标签。合肥、武汉、长沙这三座城市也同样值得关注。

- 合肥作为安徽省的省会城市，人工智能产业起点高，起步早，打造的“中国声谷”是中国首家定位于语音和人工智能领域的国家级产业基地，依托中科大、中科院、科大讯飞等平台，在“互联网+、智能终端、智慧城市、移动健康”四大产业方向进行布局，建设产业高地、完善产业链、形成产业化集中效应、打造千亿园区。近年来，合肥市也出台了多项支持人工智能创新发展的政策措施，在产业环境、技术创新、资金支持、人才服务等方面对人工智能产业给予了更多保障。

- 武汉作为全国高校最多的城市，是孕育人才的宝地、中国四大科教中心城市之一、全国三大智力密集区之一，同时也是全面创新改革试验区。在2022年二十大同期，武汉获得支持创建国家人工智能创新应用先导区，将在未来大幅提升人工智能与实体经济的深度融合，为武汉进一步培育新的经济增长点，从而实现特色化发展。

- 长沙地处华中地带，地理位置优越，四通八达。未来长沙发展潜力巨大，其经济增速在全国范围内十分醒目，这得益于长沙发达的媒体行业、完备的重工产业、知识教育产业，以及快速发展的高科技领域。因而，长沙在华中地带乃至全国内都有很大的影响力。

中国人工智能发展的城市覆盖面正从核心的地区和城市向外扩展，未来将会出现越来越多具备城市特色的人工智能示范区，为产业发展树立标杆。各地区将结合自身发展需求，在提升城市人工智能计算力基础设施水平的同时推动城市产业智能化发展，提升对外赋能能力，在相关政策的不断引导和企业的大力投入下，更多城市将深度参与人工智能创新应用，推进数字经济与实体经济的融合创新，加速产业集群转型。

最佳实践
淮海智算中心：长期布局，打造全国智算枢纽

安徽省宿州市毗邻长三角和环渤海经济区，长期布局云计算产业的发展，现已在宿州市高新区建成长三角区域建设标准最高、单体规模最大的大数据存储和处理中心，规模达十万平，被誉为“中国云都”。

推进普惠算力服务落地

为更好地推进普惠的人工智能算力服务，解决企业自己建数据中心成本高、难度大、周期长等挑战，宿州市与浪潮签署战略合作协议，共同推进淮海智算中心建设，推进人工智能产业本地的集群化发展。

- 淮海智算中心设计采用全球领先的“E级AI元脑”智算架构，部署浪潮AI服务器算力机组，通过开放多元的系统架构，支持当前先进的GPU和国产人工智能算力芯片。

- 淮海智算中心将成为国内首个算力、算法一体化的新型基础设施，运行中文智能语言算法大模型“源1.0”，通过算力的生产、聚合、调度和释放四大关键作业环节，为社会经济、产业发展和研究创新等各界提供人工智能所需算力服务、数据服务和算法服务。

- 淮海智算中心将开放支持国际国内主流的AI框架、数据集和工具，无缝对接当前国际国内成熟完善的人工智能开发和软件生态，可运行智能图像、智能语音、语言智能、机器人、自动驾驶、AI+Science等众多丰富领先的行业应用，并满足政产学研多元化人工智能场景创新的关键需求。

持续促进人工智能产业发展

淮海智算中心总体建设规模达300PFLOPS，总体投资10亿元，全面建成后智能算力性能将达30亿亿次每秒。未来，淮海智算中心将作为智算枢纽，通过领先的算力、算法基础设施，开放的技术架构，成熟丰富的生态应用，面向全国和省内提供智能算力、数据和算法服务，承接长三角、环渤海经济区、京津冀等地区的智算服务需求，加速智算产业生态在宿州落地集聚，为宿州市加速融入长三角一体化发展赢得先机，极大促进宿州市乃至安徽省的人工智能产业和数字经济的新发展。

行动建议

4

根据IDC全球针对人工智能调查结果显示，虽然人工智能应用正在稳步发展，大部分受访者表示他们已经将人工智能应用在业务中，但一些应用仍处在实验、评估和测试阶段，只有三分之一企业声称已达到成熟阶段，对于投资人工智能的企业来说，提高客户满意度、自动化决策和利用AI完成重复性工作是企业最主要的收益。IDC看到，为人工智能专门构建的IT基础设施的缺乏往往是人工智能应用无法进一步深入的原因，这一点需要行业企业、解决方案提供商以及整个人工智能产业的各个组成部分共同努力。针对以上问题，本报告提出以下建议。

4.1 对行业用户的建议

■人工智能基础设施应为企业可持续发展提供长久支撑力：

AI算力基础设施应成为企业IT基础设施建设重点。目前，大部分企业的人工智能基础设施尚未达到成熟水平，企业需要衡量的因素包括初始投资，如何实现投资收益和回报，以及如何确保基础设施能够满足业务需求和绿色低碳的要求。本报告认为，在未来几年，最值得重视的是如何提高计算能效。数据中心发展到现今的规模和体量，能耗将成为未来制约其发展的重要因素。企业可利用异构计算，提高单位能耗带来的算力，同时可积极尝试液冷技术，虽然较高的初期成本仍然是投资的最大障碍，但企业应制定更长远的目标，综合考核基础设施的性能、灵活性、易用性、成本和能效这五个方面，设立指标并考虑其影响，以选择最匹配的组合。

■ 广泛合作，充分借鉴，合力促发展：

数据处理是企业投资人工智能基础设施的另一大障碍，行业用户通常缺乏构建、培训和部署人工智能模型的时间和相关知识或能力，这为预训练的人工智能模型带来了一个新的市场。然而，与任何现成的模型一样，预训练模型也有局限性，包括模型可用性和适应性、运行模型的基础设施局限性以及内部专业知识不足。模型规模也在不断增长，这使得它们在通用基础设施上运行具有挑战性。通常，行业用户不具备进行模型二次开发的技术能力，因为这需要投入大量时间和人员成本，因此，行业用户需要更积极地参与到人工智能生态建设中，借助合作伙伴和技术供应商的能力，进一步完善人工智能在企业的应用。

4.2 对技术提供商的建议

■ 降低行业用户获得人工智能能力的门槛：

在基础设施层面，解决方案供应商必须采取一致但多管齐下的方法来应对人工智能给行业用户带来的挑战，包括广泛的成本选择以及更加完整、灵活的解决方案。技术供应商需要有能力提供一个创新、完整的解决方案，并具有足够的灵活性，允许在本地和云上集成异构、混合的计算和存储环境，并提供预测性分析，将系统健康检测标准化，最大限度减少用户在系统运维上的复杂度，降低用户应用AI的门槛。

■ 推进构建和部署模型的自动化进程：

在软件层面，目前大多数企业才刚刚开始应用机器学习，需要大量工具和能力，使他们能够轻松、快速地管理机器学习模型的实验、开发和部署。对于大多数用户来说，MLOps是一个新生的过程，它们将尝试不同的方法，以最好地处理机器学习模型的创建、维护和操作。解决方案供应商有机会在这一过程中影响和帮助其客户，随着容器化模型部署成为标准，机器学习操作的重点正转向工作流协作、模型部署和模型监控。人工智能模型的运作，主要围绕模型治理和模型生命周期管理，也越来越受到关注，解决方案提供商应关注使MLOPs可扩展的功能，包括但不限于模型部署、模型监控、模型检测和评估，以便让用户更快速地获取AI能力。

4.3 对产业发展的建议

■ 加速自身研发能力的提升：

人工智能作为一个完整的产业，其生态建设至关重要。技术提供商和行业用户应该坚持更加开放和深入的合作，共同推进人工智能应用的发展。除此之外，在学习全球先进技术的同时，需要加速自主研发的进程，无论在算法模型方面，还是算力层面，拉近与全球领先者的差距。作为算力的核心，芯片的自主研发已迫在眉睫。未来算力的需求将变得越来越多样化，聚焦在芯片研发的技术供应商应利用人工智能在中国应用多样化发展的特点，联合生态合作伙伴，提供更贴近本地用户需求的解决方案。

■ 各地区应挖掘自身特点和优势，因势利导：

从地区分布角度来看，中国人工智能应用同样具有多样化的发展趋势。各个城市和区域具有不同的环境、地理位置，以及各自的产业优势，可以此为基础，探索具有自身特色的发展路径，并给其他城市提供借鉴，共同推动人工智能产业在中国的发展。

■ 各方协作推进绿色发展：

为满足人工智能应用可持续性发展的需求，绿色低碳的AI算力基础设施是未来发展方向，例如液冷数据中心正逐渐被采用。但整体来看，其发展仍处在初期阶段，在实际建设过程中，用户面临诸多挑战，例如成本因素、数据中心改造复杂度等。产业需要进一步完善绿色数据中心基准，给用户实施提供参考，并从中受益。

关于浪潮信息

浪潮信息是全球领先的IT基础设施产品、方案和服务提供商，业务涵盖服务器、存储和网络三大领域，有8个研发中心、10个生产基地、26个分支机构，业务遍及全球120多个国家和地区。

人工智能是浪潮信息战略重点业务之一，浪潮信息是全球第二大服务器供应商，也是全球第一的 AI 服务器提供商。在中国 AI 加速计算市场占有率连续第六年超过 50%，可提供从计算平台、算法模型、管理套件、框架优化到应用加速的完整方案。浪潮信息推动 AI 领域开放计算的发展，参与制定了 OCP社区的 OAM 规范以及 ODCC 社区的 GPU 服务器规范，为不同的 AI 技术提供统一的技术标准。浪潮信息坚持“伙伴第一”的原则，不断发展元脑生态，聚合具备AI开发核心能力的左手伙伴和具备行业AI整体方案交付能力的右手伙伴，加速行业智能的构建，最终帮助用户完成业务智能转型升级。

关于 IDC

国际数据公司（IDC）是在信息技术、电信行业和消费科技领域，全球领先的专业的市场调查、咨询服务及会展活动提供商。IDC 帮助 IT 专业人士、业务主管和投资机构制定以事实为基础的技术采购决策和业务发展战略。IDC 在全球拥有超过 1100 名分析师，他们针对 110 多个国家的技术和行业发展机遇和趋势，提供全球化、区域性和本地化的专业意见。在IDC 超过 50 年的发展历史中，众多企业客户借助 IDC 的战略分析实现了其关键业务目标。IDC 是 IDG 旗下子公司，IDG 是全球领先的媒体出版，会展服务及研究咨询公司。

IDC China

IDC中国（北京）：中国北京市东城区北三环东路36号环球贸易中心E座901室
邮编：100013
+86.10.5889.1666
Twitter: @IDC
idc-community.com
www.idc.com

全文下载