中国科学技术信息研究所
INSTITUTE OF SCIENTIFIC AND TECHNICAL INFORMATION OF CHINA

人工智能计算中心发展白皮书

中国科学技术信息研究所

2020年10月

摘要
ABSTRACT

新一代人工智能（Artificial Intelligence，缩写为AI）是引领未来的战略性技术，正在与5G、大数据、物联网等领域深度融合，加速推动智能经济发展和产业数字化转型。我国高度重视人工智能发展，习近平总书记在十九大报告中指出，要“推动互联网、大数据、人工智能和实体经济深度融合”，《新一代人工智能发展规划》、《促进新一代人工智能产业发展三年行动计划(2018-2020年)》等多个国家政策陆续出台，我国逐渐形成了涵盖人工智能计算芯片、人工智能计算服务器、人工智能基础应用、人工智能行业应用及产品等较完善的人工智能产业链。

数据、算法、算力是新一代人工智能发展的三要素。以人工智能新型计算能力为代表的人工智能计算中心是新型基础设施建设的重要组成部分。随着人工智能的深入应用，算力建设分散，中小企业或科研机构难以开展复杂模型、海量数据研究的问题日益凸显，建设大规模人工智能计算中心正在成为推动人工智能产业进一步发展的关键要素。

人工智能计算中心是人工智能算力建设的重要发展方向，是涵盖了基建基础设施、硬件基础设施和软件基础设施的大规模系统工程。依托人工智能计算中心，可以打造公共算力服务平台、应用创新孵化平台、产业聚合发展平台、科研创新和人才培养平台，形成“1个人工智能计算中心 + 4个平台”的人工智能产业布局，赋能区域产业集群。

当前，人工智能计算中心仍然面临着能耗密度高、企业应用水平较低等问题，对于我国来说还面临着人工智能芯片及框架等核心技术受制于人的挑战。因此，在人工智能计算中心建设中，需要做好顶层设计、强化统筹推进，有效选择自主可控的技术路线，建立完善的运营机制，积极打造服务平台，形成以人工智能计算中心为核心支撑的人工智能产业生态，加速人工智能新兴产业创新发展，促进人工智能与传统产业深度融合，拉动区域经济转型与高质量发展。

目录
CONTENTS

第一章

人工智能计算中心概述

人工智能作为新一轮科技革命和产业变革的重要驱动力量，世界各国都在积极抢占这一战略领域制高点，力图在新一轮国际科技竞争中掌握主导权。我国于2017年印发的《新一代人工智能发展规划》中，提出了关于构建泛在、安全、高效的智能化基础设施体系的重大任务，根据智能经济、智能社会和国防建设等需求，推动智能化信息基础设施建设，提升计算中心对人工智能发展的推动作用。在国家新型基础设施建设规划中，人工智能为产业数字化转型和智能升级提供底层支撑，形成新一代信息基础设施的核心能力，将推动生产效率提升和产业结构优化。因此，人工智能基础设施的建设，既是促进产业创新能力的重要手段，也是拉动新一轮经济和社会跨越式发展的新引擎。

1.1 人工智能计算中心的概念

以人工智能新型计算能力为代表的人工智能计算中心，是人工智能基础设施建设的重要组成部分。人工智能计算中心是以基于人工智能芯片构建的人工智能计算机集群为基础，涵盖了基建基础设施（机房基建）、硬件基础设施和软件基础设施的完整系统，主要应用于人工智能深度学习模型开发、模型训练和模型推理等场景，提供从底层芯片算力释放到顶层应用使能的人工智能全栈能力。依托人工智能计算中心，行业应用将基础能力转化成人工智能技术，为各个行业领域的应用赋能。

图1. 人工智能计算中心重要组成

人工智能技术的迅速发展，对人工智能计算中心提出规模化建设需求。一方面人工智能算法愈发复杂、模型规模不断提升，图片、语音、视频等非结构化数据爆炸式增长，另一方面人工智能与5G、物联网等行业领域结合落地，使得人工智能的发展对算力的需求呈现指数级增长。人工智能计算中心的建设和发展，除了满足日益增长的算力需求，同时为大规模算法和模型的基础理论研究、实时复杂的智能化计算引擎发展、人工智能应用的商业落地、关键共性技术的研发创新等方面形成条件支撑，并将一同促进人工智能硬件、软件和智能云服务之间相互协同的生态链发展。

1.2 人工智能计算中心的关键作用

人工智能计算中心的建设，以新一代人工智能理论、技术与应用体系为支撑，发展意义重大。依托于人工智能计算中心，可以配套建设为企业提供普惠算力的公共算力服务平台、匹配本地产业特色的应用创新孵化平台、聚合产业生态的产业聚合发展平台、支撑当地科研创新和人才培养的科研创新和人才培养平台，形成“1个人工智能计算中心 + 4个平台”的人工智能产业布局，赋能本地产业集群。

2) 依托人工智能计算中心打造应用创新孵化平台

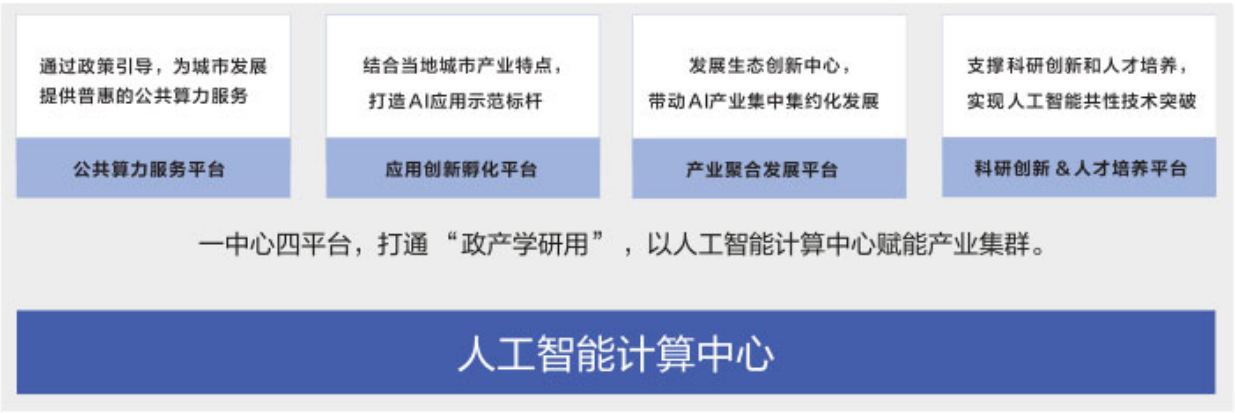


图2. 人工智能计算中心的关键作用

1) 依托人工智能计算中心打造公共算力服务平台

通过地方政府的产业政策引导，将人工智能计算中心的算力资源有序和高效的开放给当地的企业、科研机构和高校，解决当地人工智能技术发展和产业智能化转型的算力需求和服务问题。

算力是人工智能发展的重要因素之一，大规模的深度学习模型训练带来昂贵的算力成本，导致大量中小型企业和研究组织难以持续获取这种算力研发条件，成为了制约人工智能技术发展和应用的沉重负担。因此，面对广泛中小规模及初创企业、中小型研究机构的算力需求，人工智能计算中心将提供普惠的公共算力服务平台。

随着社会发展，各区域形成了具有本地特色的产业集群。依托于人工智能计算中心，可以建设有本地特色的应用创新孵化平台，与本地优势地位的产业（如制造、医疗、交通、网联车等）相结合，走一条有特色的人工智能发展道路。

可以考虑由政府或产业组织编制面向人工智能应用场景的项目机会清单，面向人工智能

企业、高校院所、科研机构进行公开发布，鼓励高校基于人工智能计算中心的算力开展人工智能竞争性和先导性应用开发和场景试验，牵引科技创新成果做商用转化、形成重大产品创新和示范应用。通过打造一批有影响力，有实际效果的应用示范项目，进一步带动当地产业的智能化升级。

3) 依托人工智能计算中心打造产业聚合发展平台。

依托人工智能计算中心，可以建设产业聚合发展平台，聚合人工智能产业链上的各类公司，包括算法公司、数据处理公司、行业集成公司等，形成完整的产业闭环，促进产业快速发展。

可以考虑建设配套园区，并建立人工智能生态创新中心等生态运营组织，通过生态创新中心进行企业交流、初创孵化、技术赋能、人才培训、技术方案对接，产业推广等活动，政府也可以针对园区出台相关配套政策，吸引和招募人工智能领域的相关企业入驻园区，从而促进和推动人工智能产业集约集聚发展。

4) 依托人工智能计算中心打造科研创新和人才培养平台

结合区域教育资源情况，鼓励高校院所联合行业龙头企业，采用产学研合作模式创建一批人工智能重点实验室、研究院等创新科研组织，基于人工智能计算中心的充沛的算力资源，围绕产业技术创新需求，开展人工智能技术研发、科技成果转化等重点工作，形成一批科技创新成果的落地和关键人才的培养。

建设高水平的人才队伍和创新团队是我国人工智能发展的重中之重。在人工智能学科建设方面，目前经教育部批准设立人工智能本科专业的高校达200余所，许多高校成立了专门的人工智能学院和人工智能研究所，建设跨学科的人才培养体系。同时鼓励企业加大对高校的投资力度，依托先进的人工智能计算平台和软件工具，支持复合型人才培养，加强与各行业领域的人工智能技术合作和探索，满足国家和区域的人工智能产业发展需求。

因此，面对日益激烈的国际科技竞争，我们需要高度重视并大力支持人工智能计算中心发展，在新一轮科技发展及产业变革中夯实基础，形成“1个人工智能计算中心 +4个平台”的人工智能产业布局，占得发展先机。

第二章

人工智能计算中心发展现状和趋势

计算是人类能力的延伸，算力的建设与社会的发展需求紧密结合，在不同历史阶段出现了超级计算中心、云计算数据中心、人工智能计算中心等不同形态的算力基础设施。人工智能计算中心是当前人工智能快速发展和应用所依托的新型算力基础设施。人工智能计算中心当前的建设现状主要为政府主导建设和头部企业自行建设，在建设的过程逐渐出现了全栈一体化、技术融合、平台化赋能等趋势。

2.1 人工智能计算中心发展现状

• 2.1.1 人工智能计算中心的演进

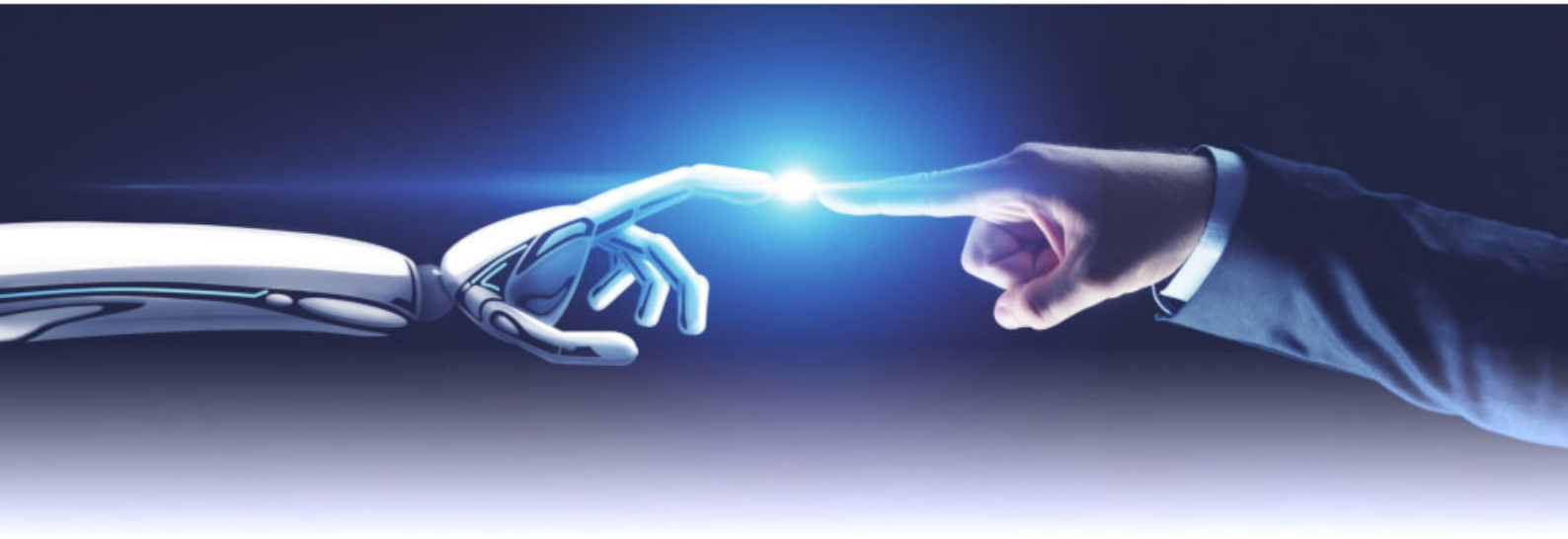
计算是人类认知世界的一种模式。从大型机到个人计算机，从智能手机到可穿戴设备，计算能力日益成为人类能力的延伸。在算力上的投入，不仅直接带来ICT 行业的增长，还对制造、交通、能源、零售、农业等诸多行业带来创新改变，推动经济增长。

算力的建设与当时社会和技术的发展需求相结合在不断演进，出现了超级计算中心、云计算数据中心和人工智能计算中心等不同形态的算力基础设施。从20世纪60年代开始，为了对重大军事研究和科学问题进行计算模拟，超级计算机和超级计算中心诞生。2000年互联网产业兴起，及2007年大数据技术和云计算技术的兴起，带动了云计算数据中心的建设。云计算数据中心面向个人或企业提供包括虚拟机计算能力、数据存储和网络传输带宽等能力，以支撑从电子商务到电子政务等方方面面的云服务。

随着2012年以来新一代人工智能技术的快速发展和突破，以深度学习计算模式为主的人工智能算力需求呈指数级增长，计算机视觉、自然语言处理等面向人工智能的处理场景越来越多，对专用定制化人工智能算力的需求大量涌现，专门的人工智能计算中心在近年来进入人们的视野。

人工智能计算中心借鉴了超级计算中心和云计算数据中心大规模并行计算和数据处理的技术架构，但以人工智能专用芯片为计算算力底座，同时软件架构和业务架构也与前两者完全不同，是当前人工智能快速发展和应用所依托的新型算力基础设施。

超级计算（Supercomputing）的历史可追溯到20世纪 60 年代，在许多情况下又被称作高性能计算（High Performance Computing，HPC），是利用并行处理和互联技术将多个计算节点连接起来，从而高效、可靠、快速地运行大规模数值求解和数值模拟应用程序的过程。超级计算机的发展经历了从向量机、大规模并行处理机到集群的体系架构的变化。



当前超级计算中心（高性能计算中心）利用超大规模的计算机集群系统，求解科学计算中的大规模并行计算问题，以双精度浮点运算性能作为主要的性能衡量指标，提供用于数值模拟的超级计算算力。随着异构计算范式的兴起，由协处理器处理其擅长的海量并行计算，可明显提高超级计算系统的计算能力，已成为当前超级计算创新发展的重点方向之一。超级计算系统包括计算节点（提供通用中央处理器或通用协处理器浮点计算能力）、存储及文件系统、高速互联网络、集群管理、资源调度等部分。超级计算中心应用的领域有模拟仿真、油气勘探、天文学、计算化学、流体力学、生物信息学、气象预报及环境模拟等，一般服务于科研院所、高校和企事业单位，以及一些国家重大科研项目等。

云计算数据中心的前身是数据中心，2000年前后，互联网迎来爆发式增长。互联网公司兴起，PC端的繁荣需要不间断的网络来传递和存储大量产生的数据，促进了数据中心的快速发展和建设。到2007年，随着大数据技术、云计算技术的兴起，数据中心利用IaaS(Infrastructure as a Service，基础设施即服务)、PaaS(Platform as a Service，平台即服务)、SaaS（Software as a service，软件即服务）实现了计算资源需求向按需定购模式的转变，演进成为云计算数据中心。

云计算数据中心，通常利用通用中央处理器能力或增加通用协处理器加速，通过分布式计算和虚拟化技术搭建服务器集群，以在Internet网络上传递、展示、计算、存储数据信息。云计算数据中心包含一整套复杂的硬件设施，包括计算资源、网络资源、存储资源、环境控制设备、监控设备、各种安全装置等，其主要特点是面向不同应用的算力需求，以免费或按需租用的方式对外提供虚拟机、裸金属、容器等多种服务，同时也可以提供人工智

能、大数据、物联网、边缘计算等新服务，服务主要面向社会公众、政府机构、各类行业企业和IT公司等。

2012年以来，以深度学习为代表的新一代人工智能技术得到快速突破和应用，并逐渐成为最重要的计算算力资源需求之一。一方面，传统的云计算数据中心及高性能计算中心，呈现出智能化服务或智能化算力的建设趋势，一定程度上提供了人工智能发展所需的算力。另一方面，以人工智能算力为主的人工智能计算中心应运而生，能够提供人工智能计算范式所需的专用算力，配合少量的通用算力以进行数据预处理和其他任务，从而能够以较低的成本提供高效的人工智能专用算力，为计算基础设施带来了新的建设方式。

人工智能计算中心基于人工智能芯片构建人工智能计算系统，主要应用于人工智能模型开发、模型训练和推理服务场景。人工智能计算中心提供软硬件全栈能力，包括基于人工智能芯片的计算资源、高速互联网络资源等硬件资源，人工智能芯片使能、人工智能计算框架、人工智能使能平台、人工智能应用市场、行业智能体等软件服务，在当前与云结合的发展趋势下，还包括云底座（提供IaaS服务、PaaS服务、运营软件等）等。

• 2.1.2 人工智能计算中心的建设现状

随着人工智能计算需求的指数级增长，人工智能算力的成本也同步高涨。MIT计算机科学家Charles Leiserson 在《Science》上发表文章《There's plenty of room at the Top: What will drive computer performance after Moore's law?》指出：深度学习正在逼近现有芯片的算力极限；计算能力提高10倍相当于三年的算法改进；算力提高的硬件、环境和金钱成本将越来越高。

人工智能可以建立超越专家的模型，也带来昂贵的算力成本，其所需的硬件设备和计算力，背后消耗的是巨额资金。在计算机视觉领域，将Efficentnet训练到需求精度，按照英伟达V100 GPU的成本估算，将需要172032美元；在自然语言处理领域，将Transformer训练到所需精度，将需要3840美元，而到了2019年的Bert模型，训练到所需精度花费将达到15360美元。

一份业内报告显示，华盛顿大学的Grover假新闻检测模型两周的训练费用约为25000美元。另据报道，著名人工智能非营利组织OpenAI花费高达460万美元训练其1750亿参数的AI模型 GPT-3语言模型，而GPT-2语言模型，每小时训练花费则达到256美元。算力门槛的提高，导致很多大学、研究机构的中小团队很难获得这种算力科研条件，同样将大量中小企业挡在门外。

因此，具备训练复杂先进模型和处理海量数据能力的人工智能计算中心属于投资较大的

信息基础设施，是包含了机房基建、硬件基础设施和软件基础设施的大规模的系统工程，当前的建设模式和现状主要为政府主导建设和头部企业自行建设。

1）政府主导建设

在国家层面，出于保持国家竞争力、带动产业发展等考量，各国政府纷纷出资或政策引导建设人工智能计算中心。

美国施行了政府主导的模式。美国能源部计划投资18亿美金进行3个E级超算（Aurora, El Capitan, Frontier）的建设，每个超算系统都包含了对人工智能计算能力及软件应用等的支持,预计在2021年到2023年分别部署在美国的三个国家实验室。2020年美国白宫科技政策办公室（OSTP）发布《美国人工智能倡议首年年度报告》，该报告概述了美国人工智能的进展，并为美国的人工智能计划提供了长期愿景。

作为美国人工智能计划的一部分，联邦机构通过酌情分配资源和资源储备来优先分配计算资源，以进行人工智能研究。例如，美国能源部与美国国立卫生研究院（NIH）的国家癌症研究所（National Cancer Institutes of NIH）合作提供了世界上最大的人工智能超级计算机，以为美国面临的紧迫的癌症挑战构建人工智能解决方案。

德国在2018年正式推出国家级人工智能战略，已资助了一批高校建设人工智能计算中心和能力中心。德国柏林工业大学在2020年1月15日宣布成立新的人工智能研究所，目标是在这一领域开展先进的科研和人才培养。研究所的主要任务包括，开展大数据、机器学习和交叉领域的尖端科研，从技术、工具和系统方面强化人工智能在科学、经济和社会中的作用，培养全球急需的人工智能专业人士。德国联邦政府将在人工智能战略框架内对该研究所追加预算，预计到2022年时，研究所将获得3200万欧元财政支持。

在我国人工智能战略和深圳“双区驱动”整体布局下，鹏城实验室和华为合作，共同研制人工智能大科学装置——鹏城云脑，以云化方案打造人工智能计算中心、面向全国的人工智能基础开源开放平台和人工智能开源开放创新生态环境。支撑粤港澳大湾区人工智能重大应用需求、提升大湾区人工智能研究基础地位与创新力和吸引全国人工智能资源、技术与人才。鹏城云脑计算能力预计达到1EFlops，存储容量预计64PB，将建设成为软硬件一体化的人工智能开源开放平台，为智能交通，气象、基因筛选等领域提供E级算力和人工智能算法支撑，逐步实现产学研用的基础公共算力平台。

2）头部企业建设

人工智能企业的创新发展和传统企业的智能化升级需要充沛的人工智能算力支撑。近年来人工智能技术领先的企业已普遍开展人工智能算力平台建设，部分龙头企业根据自身的业务特点投资人工智能专用芯片，并依托人工智能芯片建设专有集群。但大量中小规模及初创企业对人工智能算力虽有强烈诉求却无力自建，需要依靠公共算力平台获取普惠人工智能服务支撑发展。

- Google最早投入人工智能，人工智能团队拥有超过1300名研究员，拥有最大的数据库资源（大约有10到15EB的数据量）。Google于2016年发布自研TPU AI芯片，提供云、框架、芯片的全栈人工智能解决方案，TPU芯片围绕自用业务场景构建最佳性能，并逐步溢出提供AI云服务。在2019年NeurIPS大会上（全球最受瞩目的人工智能、机器学习顶级学术会议之一）以170篇论文遥遥领先。Google不仅在搜索、翻译等一系列服务中融入了人工智能技术，也在其云平台上开放了Cloud TPU和Cloud TPU Pod服务，旨在满足那些需要大量计算能力的大型人工智能项目。

- 微软持续战略投入人工智能，成立Cloud AI部门，收购大量数据和人工智能初创公司。微软在2018年5月Build大会期间宣布开发者可以接入Azure云,试用由微软基于英特尔FPGA芯片打造的低延迟深度学习计算平台Project Brainwave提供的人工智能服务。2019年7月，微软宣布向人工智能研究实验室OpenAI投资10亿美元，以共同构建一个新的Azure AI计算平台，将主要用于训练和运行更加先进的人工智能模型。

• 华为在2018年全连接大会上首次公布人工智能战略，截至目前战略投入超过3000人。华为发布了面向训练的昇腾910和面向推理的昇腾310两种人工智能芯片， 为了大幅度降低行业使用人工智能的门槛，还发布了ModelArts人工智能使能平台，从最底层的芯片开始，支持行业全场景人工智能诉求，并通过华为云EI， 为广大用户提供一站式人工智能平台服务。华为基于云化方案在公司内部部署超过10万台鲲鹏与昇腾设备和ModelArts，有力支撑了内部四大人工智能实验室的研发创新工作，打造了智能制造、自动驾驶、智能交通、智能气象、智慧城市、智能医疗、智能政务、网络运维、流程办公等解决方案，覆盖从研发、生产、办公、交付到销售的全业务场景。

• 阿里巴巴将云事业群升级为阿里云智能事业群，云和智能上升到最为重要的位置，为满足业务的需求，阿里不断深化人工智能基础设施建设，重金投入研发含光800人工智能专用芯片和超大规模机器学习平台，并建成了单日数据处理量突破 600PB 的超大计算平台，以云服务方式对外提供能力。

• 科大讯飞是科技部授予的“智能语音”国家新一代人工智能开放创新平台，科大讯飞在语音识别、自然语言理解、机器学习推理及自主学习等领域保持着国际前沿技术水平，是国内的人工智能领头羊，内部建设了高性能GPU服务器集群，以满足业务部门的研发创新工作。

• 商汤是科技部授予的“智能视觉”国家新一代人工智能开放创新平台，在WAIC 2020大会期间，商汤科技宣布，上海“新一代人工智能计算与赋能平台”临港超算中心启动动工；该算力中心占地面积近80亩，总投资金额超过50亿元人民币，一期将安置5000个等效8000W的机柜；算力中心建成并投入使用后，总算力规模将超过3700PFLOPS，可同时接入850万路视频，1天即可完成23600年时长的视频处理工作。

• 旷视是科技部授予的“图像感知”国家新一代人工智能开放创新平台。旷视在芜湖、内蒙古林格尔新区等地建设人工智能算力平台，助力当地人工智能与大数据产业的聚集和发展。

因此可见，未来人工智能企业的竞争，企业的人工智能算力支持将成为一个核心竞争因素。

中小企业或科研院所自建人工智能计算平台，面临算力不足、调优水平有限不能充分发挥硬件效率等问题，因此以政府为主导的人工智能计算平台集约化建设，为中小企业或科研院所提供公共算力服务成为趋势，一方面充分利用集约土建、电力、运维的优势，降低总体建设和运维成本，一方面可以以充沛的投资进行大规模的算力平台建设，承担人工智能领域

的国家重大战略需求、人工智能的基础共性技术攻关、人工智能领域的前瞻性基础与算法研究、当地优势产业领域的人工智能核心技术研发及应用、人工智能高端人才培养等任务。

国内外大型龙头企业持续战略投入人工智能研究和建设，以应用驱动人工智能芯片定义，建设人工智能计算中心，布局人工智能云服务能力。国外Google、Microsoft等互联网巨头率先建立人工智能计算服务能力，国内华为云EI、阿里的人工智能云服务也已对外提供服务能力。国内外龙头企业在人工智能服务能力的建设上处于齐头并进的态势，这些建设一般是企业主导，主要面向商业市场，提供商业算力服务，同时将龙头企业自身的人工智能应用能力溢出提供人工智能云服务。

在人工智能计算中心的建设过程中，逐渐形成了高质量集约化建设、提供公共人工智能算力服务的共识，出现了全栈一体化建设、多种技术融合、平台化赋能和政府与企业合建等趋势。

2.2 人工智能计算中心发展趋势

• 2.2.1 全栈一体趋势：专用人工智能芯片与软硬件协同优化提升计算效率

人工智能带来的算力需求已经远超摩尔定律。各类人工智能加速芯片适应人工智能的算法特征，进行矩阵元操作的并行化加速，或进行针对特定人工智能计算任务的精简优化，发展方兴未艾。我国人工智能芯片起步较晚，但发展较快，当前华为、寒武纪等已推出商用人工智能芯片，还不断有新的人工智能芯片出现。

不同的人工智能芯片设计与实现方式不同，当前的发展趋势是需要人工智能芯片厂商或社区开发对应的软件进行精细化匹配，以发挥硬件的最大算力。谷歌把Tensorflow与其人工智能专用芯片TPU绑定式设计协同优化；英伟达的CUDA AI开发框架将GPU与上层软件优化衔接，充分挖掘和发挥GPU的硬件潜力。华为推出MindSpore AI开发框架通过On-device特性充分发掘其AI专用芯片昇腾芯片的硬件潜力。

在2020年7月30日发布的MLPerf Training v0.7 的结果显示，人工智能专用芯片及软硬件结合的优化大幅提升了人工智能计算效率。MLPerf 是 2018 年发起的一套通用基准，用以测量和评价机器学习软硬件性能。目前有 70 多家人工智能公司和哈佛、斯坦福等 10 所研究机构参与。

在此次MLPerf基准测试中，Google基于TPUv4的人工智能计算集群打破了6项MLPerf基准测试记录。TPU专用人工智能芯片提升了针对人工智能计算特性设计的矩阵乘法效能，大幅增加存储器频宽以利于训练数据和网络模型数据吞吐，采用专门的TPU芯片内部互联技术以提升梯度参数传输效率。同

时，要使用数千块 TPU 芯片训练如此复杂的人工智能模型，Google结合TPU的硬件特性，优化了TensorFlow、JAX、Lingvo 以及 XLA 多种软件技术手段，采用了模型并行处理、大规模批次规范化、高效计算图启动以及基于树的权重初始化等方法。

另外，中科院深圳先进技术研究所提供了华为昇腾910的测试成绩，在ResNet-50测试中，相同规模的情况下，速度已经超过了英伟达的V100 GPU。在512芯片的集群规模下,华为云EI昇腾集群服务成绩为93.6秒,优于英伟达V100的120秒。这主要得益于华为云昇腾集群服务及华为云ModelArts一站式AI开发管理平台在大规模分布式训练软硬件结合优化带来的加速比上的优势,其在512和1024芯片下可达到80%以上的加速比,分布式加速比远超英伟达和谷歌。

由此可见，人工智能计算范式定义人工智能芯片，软硬件协同提升计算效率正成为当前人工智能计算发展的新特征，人工智能专用算力与配套软件全栈一体建设成为必然。

• 2.2.2 技术融合趋势：超级计算与人工智能融合，云与人工智能融合

超级计算的业务需要大量人工智能算力，超级计算计算中心拥有支撑人工智能的能力已经是一种趋势。人工智能正在改变传统超级计算的求解方法，将人工智能技术融入超级计算系统，可以提高准确性、加快时间并降低成本。在应用驱动下，人工智能算法在医疗诊断、天文探测、地震预测等领域快速发展。从近年来超级计算应用最高奖项“Gordon Bell Prize（戈登贝尔奖）”，可以看出人工智能技术在超级计算领域的渗透。“戈登贝尔奖”是国际上高性能计算应用领域的最高学术奖项，被称为“超算领域的诺贝尔奖”，是考察全球顶尖的高效、创新、面向解决重大问题的超级计算应用，在2018年入围“戈登贝尔奖”决赛的6个应用中，有半数以上属于超级计算与人工智能融合的应用范例，比如E级深度学习气候分析、治疗阿片类药物成瘾症、时间演化地震城市问题求解器、深度学习电子显微图像处理等。

在超级计算和人工智能融合发展的趋势下，全球超算出现了大量的人工智能算力建设。从2020年6月Top500超算榜单来看，超算系统的人工智能算力建设已成核心需求。越来越多的系统针对人工智能设计了低精度算术逻辑单元，以支撑人工智能计算能力需求，尤其榜单前十名都有支撑人工智能计算的能力。

2018年美国能源部橡树岭国家实验室宣布启用了当时世界上最快的超级计算机Summit，这台计算机价值2.8亿美元，由IBM总包设计建设，采用了9216颗IBM Power9处理器和27648颗英伟达 Volta GPU加速器。Summit是史上第一台既支持传统计算、也支持运行人工智能应用程序的超级计算机。Summit的任务规划包括了“疾病和成瘾”项目，研究人员将使用人工智能来识别人类蛋白质和细胞系统的功能和进化模式。这些模式能帮助人类更好地了解阿尔茨海默病、心脏病或成瘾，进而助力药物发现过程。

2020年日本发布了投入10亿美元打造的“富岳”（Fugaku）超级计算机，这是史上第一台基于ARM架构的冠军超算，由396个机架、152,064个节点组成。“富岳”采用了全CPU同构架构，搭载Fujitsu的48-core A64FX SoC ARM处理器，采用SVE扩展指令集，双精度浮点性能415.5 PFlop/s，人工智能计算性能达到1.45EFlop/s (FP16&INT8精度)。虽然富岳系统硬件定制化程度较高，但全CPU同构架构使其可以大量复用ARM生态软件和应用。“富岳”不仅在Top500的Linpack（HPL）双精度浮点计算排名第一，同时在HPL-AI（侧重人工智能低精度计算性能）、Graph500（侧重于大数据分析等领域计算能力）、HPCG（高性能共轭梯度计算）等榜单中均排名榜首，整体表现出色。

随着云计算的发展和成熟，以云化方案构建人工智能服务，对用户提供统一架构、统一服务和统一API，向用户屏蔽复杂的人工智能技术细节，降低了人工智能服务的使用门槛。目前云提供商如华为云、AWS是这一趋势的主要推动者，纷纷推出云上高性能人工智能计算和人工智能使能平台服务，单个用户可以创建数千处理器规模的高性能人工智能计算资源满足高效人工智能开发。云平台同时也带来了人工智能计算中心运营模式的改变，通过云上租户粒度的安全隔离、完善的运维运营系统，人工智能计算中心可以为不同用户提供安全可靠、按需使用、弹性伸缩、有服务等级保障的自助式服务。云化计算中心提供裸金属服务器、虚拟机、容器等多样化的算力资源和人工智能使能平台服务，人工智能服务与云上大数据、物联网、边缘计算等服务的相互协同，满足新型应用场景综合复杂多层次的计算需求。

随着人工智能的快速发展，超级计算、云计算和人工智能技术不断融合发展，人工智能

为超级计算提供新的计算求解方法，使其可以利用长期积累的大量观测数据，云计算一方面为人工智能提供算力和新的运营和赋能方式，一方面将人工智能能力通过云与边缘计算、物联网等结合，推动人工智能在云边端全场景的应用。

• 2.2.3 平台赋能趋势：人工智能计算中心赋能企业，形成算力生态

平台是可跨情境应用的资源和能力的集合，人工智能计算中心作为集超级算力和海量数据的超级大脑，已经呈现平台化发展的趋势。具备强大软硬件能力的核心企业集聚研发能力、生产经验和产业资源，在人工智能计算中心搭建基础应用平台，并依托平台的共享输出上层应用使能能力，对平台上的小型人工智能企业和欠缺人工智能能力的传统企业进行赋能。人工智能计算中心将成为人工智能核心企业和大量初创企业能力输出的主要方式，如通过平台开放接口的方式输出龙头企业的算法能力，资源、数据支撑、运营辅导和模式优化等。人工智能计算中心逐渐构建起人工智能的生态创新架构，助力各类架构缔结产业联盟，聚拢上下游资源，吸纳高校和科研机构，提供基础研究课题和依托平台，推动专家合作、举办生态研讨等人才培养活动，为产业发展提供人才支持，最终形成算力生态，大量生态互补者协同推进人工智能产业的开发与应用。

当前如华为云的全栈全场景人工智能解决方案，覆盖10多个行业和600多个人工智能项目实践,可提供210多个解决方案。华为云EI企业智能依托华为云，将华为在人工智能领域的技术积累以云服务的方式开放赋能周边企业快速地使用人工智能，助力企业数字化转型。华为云还将华为2012实验室等在内的人工智能前沿的算法和理论研究,例如:小样本训练能力、半监督学习能力、神经网络自动搜索能力等,逐步产品化和平台化,并开放给开发者。

综上，人工智能计算中心建设之后必然会形成上层应用使能能力，溢出对周边企业进行赋能，形成科研和人才培养的算力生态，应用使能能力也成为人工智能计算中心建设的核心指标和核心竞争力之一。

第三章

人工智能计算中心总体架构与关键技术

目前，全球人工智能产业的生态系统正逐步成型。依据产业链上下游关系，如图3人工智能计算中心总体架构所示，这里将人工智能计算中心的总体架构划分为基建基础设施（机房基建）层、硬件基础设施层和软件基础设施层，在人工智能计算中心之上，是行业应用层。

基建基础建设层包括土建、电气和风水火电建设等底层设施，为人工智能计算中心提供空间、水电、散热等基本条件。行业应用层是人工智能产业的核心，将基础能力转化成人工智能技术，如计算机视觉、智能语音、自然语言处理等应用算法研发，广泛应用到多个不同的应用领域。因基建基础设施和行业应用不属于人工智能计算中心的基本属性，故不在这里展开。

在硬件基础设施、软件基础设施的技术领域中，已产生具有代表性的技术路线和厂商，如图4人工智能计算中心关键技术与代表企业所示，不同技术路线呈现并行发展态势。

图3. 人工智能计算中心总体架构

图4. 人工智能计算中心关键技术与代表企业

3.1 硬件基础设施

硬件基础设施层是人工智能计算中心的核心基础，由人工智能计算子系统、存储子系统、网络互联子系统组成。人工智能计算子系统主要提供硬件算力，由人工智能芯片、基于人工智能芯片的服务器与芯片间和服务器间互联网络构成。存储子系统、网络互联子系统围绕计算子系统提供数据存储传输、人工智能网络模型参数传输更新等功能。其中，人工智能芯片是人工智能硬件基础设施中人工智能算力最重要的承载。

• 3.1.1 人工智能芯片

依据承担的功能，人工智能芯片可划分为训练和推理芯片。训练芯片涉及到海量数据和大规模计算，对算法、精度、处理能力要求非常高，当前仅适合在中心端部署。推理芯片更加注重综合能力，包括算力能耗、时延、成本等因素，除可以在中心端部署外，还可以在边缘端或终端侧部署。 如图4人工智能计算中心关键技术与代表企业所示，目前GPU（通用型）、NPU（专用型）、FPGA（半定制化）、ASIC（全定制化）是人工智能芯片行业的主流技术路线。不同类型芯片各具优势，在不同领域已实现实际落地应用。

GPU（Graphics Processing Unit）

的设计和应用均已成熟，占领人工智能芯片的主要市场份额。GPU擅长大规模并行运算，可并行处理海量信息。在全球范围内，英伟达形成寡头垄断，占GPU市场份额的70%-80%。由于国外GPU巨头具有丰富的芯片设计经验和技术沉淀，同时又具有强大的资金实力，在全球建立了良好的GPU生态。

FPGA（Field Programmable Gate Array）芯片具有可硬件编程、配置高灵活性和低能耗等优点。FPGA技术壁垒高，市场呈双寡头垄断：赛灵思（Xilinx）和英特尔（Intel）合计占市场份额近90%。其他厂商处于追赶阶段，技术和产品成熟度有一定差距。

ASIC（Application Specific Integrated Circuits）是面向特定需求（这里是人工智能的计算范式）设计的定制芯片。它的编程方式是直接在物理硬件（门电路）上搭建电路，由于不需要取指令和译码，每个时间单位都能专注于数据处理和传输。由于需要底层硬件编程，ASIC需要大量的物理设计、时间、资金及验证，但在量产后，其性能、能耗、成本和可靠性都优于GPU和FPGA。目前，ASIC芯片市场竞争格局比较分散。

NPU（Neural network Processing Unit），在电路层模拟人类神经元和突触，并且用深度学习指令集直接处理大规模的神经元和突触，一条指令完成一组神经元的处理。NPU通过突触权重实现存储和计算一体化，从而大大提高了运行效率。在海外，谷歌TPU是NPU先行者，国内ICT巨头（如华为昇腾芯片）、初创芯片企业（如寒武纪思元芯片），也有所建树。

下表对多种AI芯片做一简单比较：

AI芯片	代表厂商	可编程性	训练芯片	推理芯片
GPU	英伟达	软件编程	Tesla A100	Tesla T4
FPGA	赛灵思	硬件编程	通过硬件编程具备一定训练能力	通过硬件编程具备一定推理能力
ASIC	———	不可编程	一般不具备训练能力	部署在终端，如无人驾驶和智能摄像头中
NPU	华为，寒武纪，谷歌，阿里	软件编程	华为昇腾910，谷歌TPUv4	华为昇腾310，寒武纪思元270，阿里含光800

表1 典型AI芯片技术路线与代表

• 3.1.2 硬件基础设施

面对海量的数据训练计算，常规的单机计算模式已经不能支撑。所以当前的人工智能计算模式必须将巨大的计算任务分成小的单机服务器可以承受的计算任务，以并行计算的集群架构提供基础的计算算力，这个集群架构就是人工智能计算子系统，如图5硬件基础设施架构图所示。每台单机服务器均搭载人工智能芯片，以人工智能芯片为人工智能算力的主要承载，在芯片和服务器之间通过互联网络传递人工智能网络模型的梯度参数更新等数据。围绕计算服务器组成的计算子系统，还配置有存储子系统、网络互联子系统，分别用于存储人工智能网络模型的训练数据和参数和传输相关数据。

性能强大、均衡和可靠节能的人工智能硬件基础设施应满足如下要求：

1）高计算密度：采用适合张量计算的人工智能芯片架构，提供高人工智能算力和能效比。

2）高速互联技术：集成多级芯片高速互联系统，提升整个集群的扩展性。

3）高度集成化设计，简化交付部署流程，节省机房空间

4）模块化和冗余设计：主要部件采用模块化集成设计，易于维护扩容，多冗余设计满足可靠性要求。

5）高效率液冷散热系统：采用部件级液冷技术，打造绿色节能的集群系统。

AI计算子系统的集群架构一般由高密度、集成化机柜式设计的集群基础单元组成，每个集群基础单元由若干AI服务器组成，一般提供不少于64颗人工智能芯片的AI算力，能够支持约40KW的散热功耗，实现低PUE数据中心能源效率。典型的AI集群机柜基本单元如华为Atlas 900 PoD（型号9000），英伟达DGX POD等。

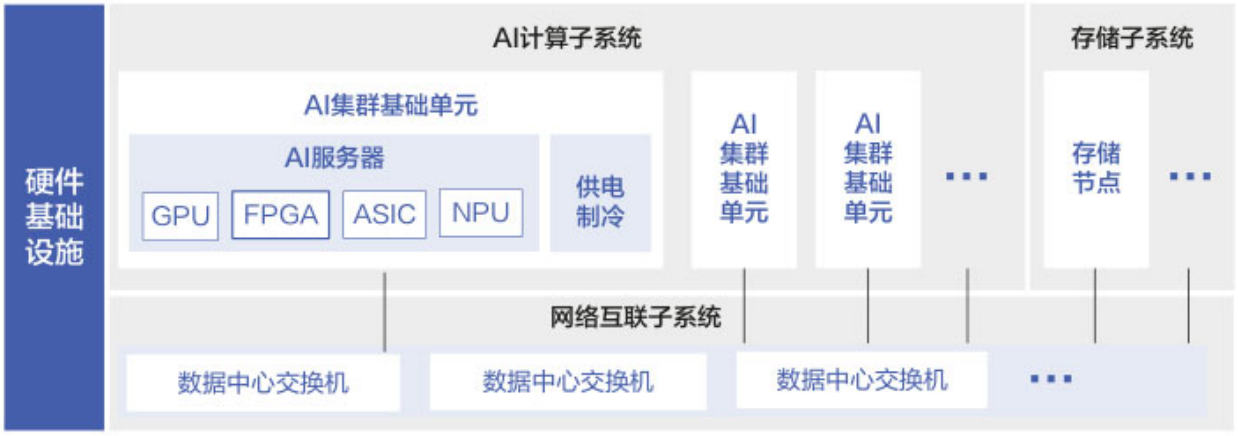
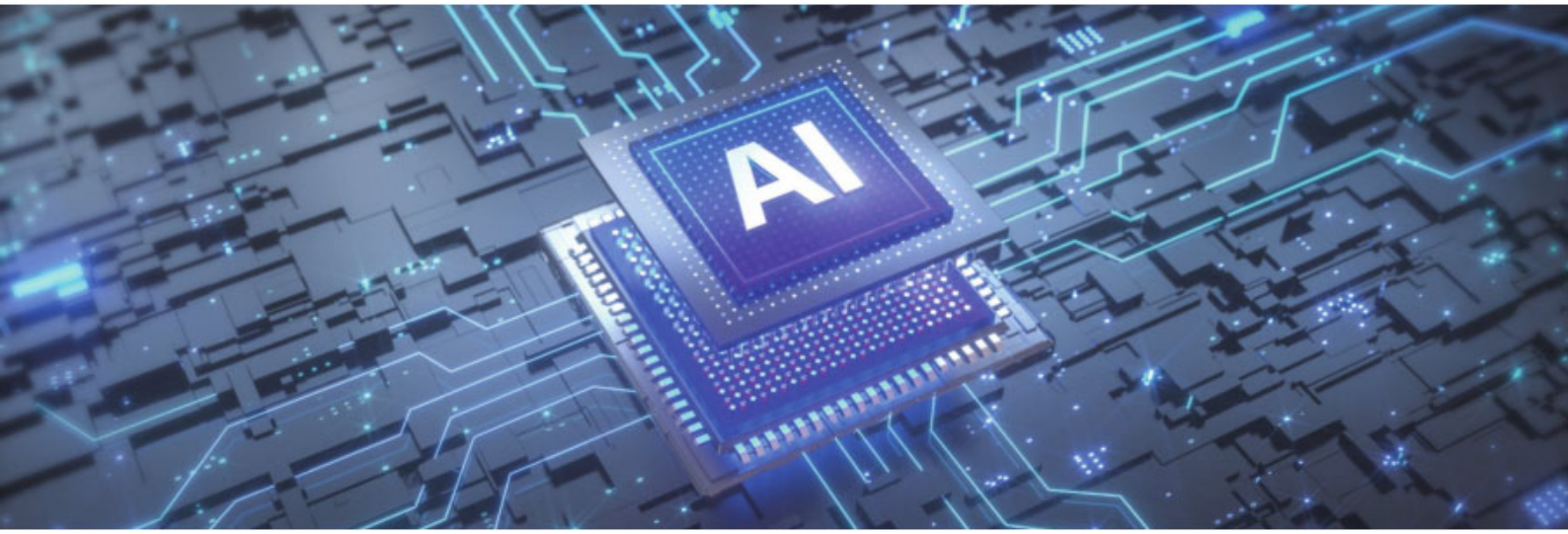


图5. 硬件基础设施架构图

存储子系统提供高性能、高可靠、高扩展性和易备份的分布式存储。存储子系统部署存储节点，提供对象存储、块存储等存储服务，为人工智能训练平台提供高吞吐，大带宽的样本原始数据。

网络互联子系统为整个AI硬件基础设施的子系统间提供互联互通。AI计算子系统、存储子系统通过网络互联子系统的数据中心交换机实现互联互通，搭建全互联无阻塞参数同步网络和数据传输网络。



3.2 软件基础设施

如图6软件基础设施架构图所示，软件基础设施层包含基础软件（AI系统软件包括AI开发框架和芯片使能软件，以及云平台）、AI使能软件、行业算法、AI市场。基础软件中，芯片使能软件驱动AI芯片，提供深度学习软件加速库（算子）的集合，AI开发框架封装了如卷积等基本操作，提供人工智能网络模型开发环境；基础软件还包括云平台，对计算、存储、网络资源进行统一调度和管理，提供统一算力支持。AI使能软件支持海量作业的自动调度、大规模分布式训练，对AI计算子系统的算力资源进行统一管理、调度和实时分配，提供算子开发研究、神经网络开发研究、全流程AI开发能力，帮助AI开发者和科研人员高效完成算子开发、算法开发、数据处理、模型训练和模型部署等开发活动。行业算法通过行业知识的积累，预置行业经验，实现更快更高效的赋能行业。AI市场支持数据和AI模型的有效流动和共享。

图6. 软件基础设施架构图

- **3.2.1 基础软件**

1）芯片使能软件

AI芯片使能软件是人工智能软件加速库（算子）的集合，这些库建立在AI芯片的驱动层之上，提供对于深度学习必不可少的计算优化功能。各大芯片厂商都推出了针对自己芯片优化的使能库。基于GPU和NPU的AI芯片路线，CUDA和CANN分别是其芯片使能软件的代表，也可能会成为未来并列的两大平台。

CUDA（Compute Unified Device Architecture）是英伟达推出的完整的GPGPU（General-purpose computing on graphics processing units，通用GPU计算）芯片使能软件，提供了对英伟达GPU硬件的直接访问接口，采用C语言作为编程语言提供大量的高性能计算指令开发能力。从CUDA体系结构的组成来说，包含了三个部分：开发库、运行环境和驱动。

CANN（Compute Architecture for Neural Networks)是华为公司针对AI场景推出的异构计算架构，通过提供多层次的编程接口，支持用户快速构建基于昇腾平台的AI应用和业务。CANN的代码是普适的，对端、边、云全场景下的训练、推理硬件没有特定依赖，为用户提供了丰富的算子库和难易结合的编程方法，开发一套代码，就能在各种终端硬件上复用，性能发挥到最好。CANN可以通过AscendCL模块提供Device管理、Context管理、Stream管理、内存管理、模型加载与执行、算子加载与执行、媒体数据处理等C++ API库供用户开发深度神经网络应用，用户可以用任何包括TensorFlow、Pytorch等第三方框架调用这些API库，而无需关心计算资源优化的问题。

2）AI框架及主流框架简介

为了把卷积运算、激活函数、损失函数计算、优化器使用等标准化，避免重复造轮子，AI框架应运而生。

早期的AI框架可以追溯到2007年由蒙特利尔大学开发的Theano和加州大学伯克利分校的贾扬清在2014年开发的Caffe。随后，各大互联网公司开始进行AI框架的开发。下面介绍主流的几种AI框架。

MindSpore

MindSpore是端边云全场景按需协同的华为自研AI框架，于2020年3月全面开源。MindSpore提供全场景统一API，为全场景AI的模型开发、模型运行、模型部署提供端到端能力。

MindSpore是一种全新的深度学习计算框架，旨在实现易开发、高效执行、全场景覆盖三大目标。为了实现易开发的目标，MindSpore采用基于源码转换（Source Code Transformation，SCT）的自动微分（Automatic Differentiation，AD）机制，该机制可以用控制流表示复杂的组合。函数被转换成函数中间表达（Intermediate Representation，IR），中间表达构造出一个能够在不同设备上解析和执行的计算图。在执行前，计算图上应用了多种软硬件协同优化技术，以提升端、边、云等不同场景下的性能和效率。

MindSpore支持动态图，更易于检查运行模式。由于采用了基于源码转换的自动微分机制，所以动态图和静态图之间的模式切换非常简单。为了在大型数据集上有效训练大模型，通过高级手动配置策略，MindSpore可以支持数据并行、模型并行和混合并行训练，具有很强的灵活性。此外，MindSpore还有“自动并行”能力，它通过在庞大的策略空间中进行高效搜索来找到一种快速的并行策略。

TensorFlow

Google公司在2015年发布了著名的TensorFlow。TensorFlow是一个使用数据流图进行数值计算的开放源代码软件库，最初由Jeff Dean领头的谷歌大脑团队基于谷歌内部第一代深度学习系统Distbelief改进而来的通用计算框架。相比Distbelief而言，TensorFlow的编程模型更加灵活、性能更好、支持更宽广的异构平台型。在激烈的商业竞争中，更快的训练速度是人工企业的核心竞争力。而分布式TensorFlow意味着它能够真正大规模进入到人工智能产业中，产生实质的影响。

有Google的背书,并将TensorFlow免费开源，吸引了众多的开发者，TensorFlow社区更新的速度快,硬件平台支持广泛,文档齐备,让很多刚开始学习深度学习的开发者不至于重复造轮子,因此又吸引了更多的使用者。从而形成良好互动的开发生态。

TensorFlow框架是目前深度学习框架中社区活跃度最高、发展最快速的一款深度学习工具，有大量活跃的社区开发者和谷歌的全力支持。截至2020年，TensorFlow在流行程

度上处于遥遥领先的位置。

Pytorch

PyTorch是原本的Torch框架基于Python的实现，主要由Facebook的人工智能研究团队(FAIR)开发。Pytorch的1.0稳定版在2018年底才刚刚发布，是一个相对比较年轻的框架，但推出后受众越来越广泛，近来和TensorFlow的差距在不断缩短。Pytorch支持动态计算图，和TensorFlow基于静态计算图的模型部署方式相比，在定义网络的操作上更简单，在调用函数时更灵活，在修改网络的某一层或者某些变量时也比较方便。甚至2019年初TensorFlow 2.0的动态图新特性的推出都被很多人认为是借鉴了PyTorch的优点。

PaddlePaddle（飞桨）

飞桨产业级深度学习开源平台以百度多年的深度学习技术研究和产业应用为基础，集深度学习核心训练和预测框架、基础模型库、端到端开发套件、工具组件和服务平台于一体，2016年正式开源，是中国开源开放、技术领先、功能完备的产业级深度学习平台。PaddlePaddle 包括核心框架、工具组件和服务平台三大部分。在核心框架层面，可以为开发者提供开发、训练和预测三大能力，在此之上，百度提供包括视觉、自然语言处理等在内的丰富模型，通过模块化的方式提供给使用者。

3）云平台

云平台使用云计算技术重构AI计算基础设施，对计算、存储、网络资源进行统一调度和管理，提供裸金属服务器、虚拟机、容器等多样化的算力资源。除人工智能外，也可以对大数据、HPC等业务提供统一算力支持，云平台同时支持AI、大数据、HPC业务等多种主流框架，既支持传统HPC的集群作业编排调度方式，也支持AI、大数据等新兴业务所需要的云原生作业编排调度方式，为人工智能、大数据、HPC等多种先进技术的高度协同和深度融合奠定基础。下面介绍几家主流的云平台服务提供商。

华为云是华为的云服务品牌，华为云提供除基本的计算、存储、网络等服务外，还可支持弹性伸缩、负载均衡、云监控、云备份、边缘计算、数据库、区块链、安全等80余种丰富的云服务，采用虚拟私有云等云计算先进技术对算力、网络、数据等资源进行全方位的完全隔离。

阿里云是全球领先的云计算及人工智能科技公司，以在线公共服务的方式，提供安全、可靠的计算和数据处理能力，让计算和人工智能成为普惠科技。

腾讯云是腾讯打造的云计算品牌，为全球客户提供领先的云计算、大数据、人工智能服务，以及定制化行业解决方案。

• 3.2.2 使能软件

人工智能计算中心面向的是大规模分布式模型训练、全流程人工智能应用支撑，需要对大规模的算力资源进行管理和调度。使能软件统一资源调度，基于硬件基础设施的组网特点实现对算力资源的统一管理、调度和监控，进行细粒度的资源实时分配，支持海量任务的智能自动调度、任务管理、数据加载和预处理，支持大规模AI计算场景，并能够提供丰富的人工智能场景应用和API服务，使能用户在该平台上的一站式人工智能开发和应用部署。软件API服务包括如智能语音语类服务和计算机视觉服务等：智能语音语类服务主要提供语音语义相关的在线服务，可包括语音识别、语音合成、声纹识别、语音听转写等。计算机视觉类服务主要提供物体检测、人脸识别、人脸检测、图像识别、光学字符识别（Optical Character Recognition，OCR）识别、智能鉴黄等服务。下面介绍一些典型的使能软件。

华为ModelArts

ModelArts是企业级的一站式AI使能平台，不仅支持对大规模AI硬件资源的统一任务调度，也为机器学习与深度学习提供海量数据预处理及半自动化标注、大规模分布式训练及加速、自动化模型生成，产业联邦训练，及端-边-云协同的模型按需部署能力，帮助用户实现模型的快速创建和部署，沉淀行业知识和管理全周期AI工作流。ModelArts支持华为全栈AI使能能力，并基于华为AI生态持续对客户进行赋能和技术更新，用户可以使用ModelArts进行人工智能算子开发、模型开发、模型训练、超大模型分布式训练、模型推理等AI业务，内嵌MindSpore 、 Tensor-Flow和Pytorch等主流 AI开发框架，底层兼容基于鲲鹏通用处理器和昇腾AI芯片的国产服务器。

ModelArts支持自动机器学习，可根据用

户标注数据全自动进行模型设计、参数调优、模型训练、模型压缩和模型部署全流程，无需代码编写和模型开发经验。模型部署上线后基于内置智能算法监控模型推理效果，自动发现推理难例进行智能诊断，提高AI模型的应用效果。

AWS SageMaker

Amazon SageMaker是AWS于2017年11月推出的云机器学习平台。SageMaker使开发人员能够在云中创建，训练和部署机器学习模型。SageMaker还使开发人员能够在嵌入式系统和边缘设备上部署ML模型。

阿里 PAI

阿里云PAI（platform of AI）是阿里云的人工智能PaaS平台，为传统机器学习和深度学习提供从数据处理、模型训练、服务器部署到预测的一站式服务。

百度 AI Studio

AI Studio是基于百度深度学习平台飞桨的人工智能使能平台，提供在线编程环境、免费GPU算力、海量开源算法和开放数据，帮助开发者快速创建和部署模型。

第四范式 SageEE

由第四范式开发的拥有自主知识产权的企业级人工智能平台，采用大规模分布式架构实现高效的离线计算和实时计算，能力覆盖从数据处理、模型调研、应用构建、应用上线到AI治理全流程，为企业和开发者提供企业级的AI平台级解决方案。

3.3 人工智能计算中心的关键发展问题

人工智能计算技术作为飞速发展的新兴技术，仍在不断的演进和变化，人工智能计算中心也在建设过程中不断面临新的形势和新的挑战，如人工智能专用芯片和人工智能框架发展协同问题，大规模建设带来的高能耗问题和赋能企业应用问题等。

• 3.3.1 人工智能芯片及框架技术独立发展，互相适配难度大

在实际工程应用中，人工智能算法可选择多种AI开发框架实现，训练和开发人工智能模型也可有多种AI芯片选项，这就开发者带来了不小的挑战。

当前人工智能芯片和应用场景、目标算法有着紧密的结合。人工智能芯片为了满足应用场景与目标算法的不同需求，而设计发展出了不同的架构与实现方式。对于应用场景的要求，芯片需要考虑计算实时性、功耗、重构开销、面积成本甚至计算精度，而目标算法决定了芯片所需要支持的算法模型，是否需要足够的通用性以解决包含经典机器学习算法以及新兴深度学习模型在内的推理算法，是否需要进行支持在线模型训练，类似的设计选择可能导致

芯片的架构走向不同的实现方式，例如面向语音识别的芯片和主要用于视觉处理的架构可能采用截然不同的结构甚至指令集，前者可能只需要支持RNN模型而选择矩阵乘阵列的设计架构而后者则可能选择传统脉动阵列结构实现。

对于相应的AI框架部分，算法开发人员可能会使用多个框架（TensorFlow、Caffe、MXNet等）进行开发，而每个框架中的工作负载都以自己独特的方式表示和执行，即使是一个简单的卷积（Convolution）操作，不同框架都可能有不同的方式定义。

因此，多种AI芯片与多种框架的适配需要AI软件底层为每个不同操作做适配，不但会导致软件设计的臃肿，影响算法执行的效率，而且为不同底层硬件和不同框架设计高效率的操作难度很高，为开发者带来大量工作量。

解决这个问题的可行办法之一是将基于不同硬件不同框架的模型文件编译成统一的、硬件能够识别的控制指令集。由AI框架或单独的中间表示层软件读取网络模型文件，找到相应的适配工具解析出计算流图，将不同框架下的计算流图转换为统一的格式对每层的操作（如卷积、激活、池化、残差等）类型，及网络参数、依赖关系、输入输出大小等参数，编译成硬件能够识别的控制指令集，写入可执行文件，并将它们调度到不同的硬件设备上。对此业界也有一些先期的技术尝试。

总体而言，AI芯片与AI框架虽然互相有一定适配，但各自有独立发展道路，不同AI芯片和不同AI框架的互相协同仍需大量底层开发工作，对此现在还没有成熟统一的解决方案。

• 3.3.2 人工智能算力能耗巨大，总体拥有成本高

功耗是影响算力水平发展的重大因素之一，在单位算力功耗呈现出逐年递减，且正向着极限逼近的态势下，算力的整体量级却仍然随着其广泛的应用而持续大规模增长。以中国为例，随着人工智能、物联网、区块链等技术的发展，中国算力中心的总用电量连续8年增速超过12%。至2018年，中国算力中心总用电量为1,608.89 亿千瓦时，占中国全社会用电量的2.35%，已经显著超过上海市2018年全社会用电量（1,567 亿千瓦时）。面向未来，算力将在信息技术产业的大力发展下继续增长，2023 年将较2019 年增长66%，年均增长率将达到10.64%，功耗总量将进一步提高。

因此优化高能耗的算力中心的PUE（Power Usage Effectiveness，电能使用效率）水平将成为算力发展一大核心挑战。算力中心PUE 优化的一大核心方向在于建立更多超大型（>10,000个标准机架可共置至少100,000台服务器）和大型算力中心（>3,000 个标准机架可共置放至少30,000台服务器），从而带动整体算力中心PUE 值优化。在中国，超大型算力中心的PUE

水平（1.5）较整体平均值（2.2）高出31%，通过测算，当前中国算力中心每降低0.1 PUE，可节省总发电量73 亿千瓦时，足够为上海全市提供17 天电力支持；倘若可将算力中心PUE 整体降至大型算力中心水平，则可节省373 亿千瓦时，足够为上海全市提供近3 个月的电力支持，同时相当于减少二氧化碳排放量约3,000 万吨，造林约8万公顷。根据Google对不同类型算力中心的TCO（Total Cost of Ownership，TCO）调研，中小型算力中心的TCO在12-25美金/瓦，而大型算力中心的TCO在8-10美金/瓦，大型算力中心相较中小型算力中心可以降低100%以上的TCO。

人工智能计算中心PUE优化的另一个核心方向是硬件基础设施的散热优化。如将AI计算子系统中的AI集群基础单元以液冷式整机柜交付，支持全液冷部署。液冷智能机柜采用整体液体冷却方式，柜内所有设备全部通过液冷模块和风液换热器散热。相比于空气，水的比热容更大，能够带走更多的热量，效率更高，更节能。液冷解决方案具有绿色节能、高可靠、高集成、易维护等特点。以鹏城实验室云脑为例，其AI集群基础单元液冷智能机柜通过高可靠和高效设计的一体成型液冷冷板，对服务器中的AI处理器、CPU等主要发热部件进行直接液冷冷却，其余通过风液换热器进行间接冷却。通过直接液冷和间接液冷的相结合，实现整柜级的整体液冷冷却，支撑鹏城云脑高性能人工智能计算业务的实现，促进深圳及粤港澳大湾区人工智能产业的可持续发展。

因此，如何提高设备利用率、节能减排、落实可持续发展的效果，是人工智能计算中心未来需要解决的问题。

3.3.3 企业应用水平参差不齐，基础数据集不足

传统产业的智能化转型困难，AI算力平台使能行业应用门槛高，企业应用水平参差不齐。当前各传统行业人员对人工智能概念的理解和算法技术的掌握难以支撑其智能化改造升

级。在业务开展初期，传统企业内部的算法团队基于自身能力针对应用开发做规划，由此导致了一种小作坊式的生产局面。作坊式生产方式在早期有其积极的一面，能够保证创新的灵活性，但是在业务发展中后期，这种生产模式的局限变得明显：与业界先进水平脱节，重复造轮子，无法形成规模化效应，整体投入产出的效益大打折扣，导致企业智能化升级难以为继。另一方面，人工智能龙头企业或某些初创企业的先进算法、模型接触不到大量用户，难以形成快速迭代的发展势头，不能满足实际业务需求。AI算法、模型等AI生产力要素无法充分流动，使其得不到有效的利用。因此需要一个低门槛、开放、端到端的人工智能使能平台，方便传统企业数据科学家和算法工程师快速开始利用平台的资源——不仅包括计算力，还包括人工智能龙头企业或初创企业提供的先进的行业应用算法模型等——执行数据准备、模型开发、模型训练、评估和预测等任务，方便地将人工智能能力转化为服务API，加速与业务应用的集成。

算据方面，由于大数据技术的出现和使用时间还不长，各类基础数据不论从数量上还是从质量上来看，都尚需要较长时间的积累。一方面，某些关键领域和学术数据集还严重不足。另一方面，已有规模化的基础数据集不仅数据质量良莠不齐，而且基本上由少数几家巨头或政府所掌握，鉴于监管和竞争等因素，无法实现有效流动。

中国移动互联网的蓬勃发展和巨大的人口基数，为互联网巨头的AI研究贡献了高质量的庞大数据集，目前全国从事数据标注业务的公司约有几百家，全职的数据标注从业者有约20万人，兼职数据标注从业者有约100万人。但中国尚缺少高质量的开源训练数据集，同时也欠缺如公共数据资源库、标准测试数据集等人工智能基础数据平台，大部分中国AI学者的研究其实用的都是海外的数据集更多。一些AI研究学者苦心整理的公开数据集排名，来自中国的数据集几乎没有，一定程度上体现了中国在开源数据集领域与国外的差距。根据德勤公司的调查，16%的IT主管将数据问题列为与人工智能相关的最大挑战，比任何其他问题都要高，39%的受访者将数据列入前三个令人担忧的方面。

因此，促进算法、算据等AI生产力要素的流动，使其得到充分利用，是有效发挥人工智能算力中心作用的必要前提。

人工智能中心的发展，面临着AI芯片和AI框架各自独立发展、协同适配难度大，AI算力能耗密度大、拥有成本高，以及企业应用能力不足、基础数据缺乏等涵盖基建基础设施、硬件基础设施和软件基础设施不同层次的问题。尤其对于我国来说，还有AI芯片及框架等核心技术仍落后，亟需发展自主可控核心技术的挑战。

AI芯片是人工智能算力的核心承载。由于创新难度大、技术和资金壁垒高等特点，AI芯片市场主要被欧美等少数国际巨头垄断。受限于技术积累时间不长与研发投入的不足，国内在芯片领域相对薄弱。在AI芯片领域，国际科技巨头芯片已基本构建产业生态，而中国企业刚刚起步，但发展较快。目前国内AI应用所采用的AI芯片市场份额95%以上被美国英伟达、AMD等占据，对我国未来人工智能战略造成巨大的业务连续性和“卡脖子”风险。

AI开发框架层实现算法的模块化封装，为应用开发提供集成软件工具包。AI框架生态的核心，是通过使用者和贡献者之间的良好互动和规模化效应，形成现实意义的标准体系和产业生态，进而占据人工智能核心的主导地位。当前得到广泛应用的深度学习框架如TensorFlow、Pytorch等都非国产，国内自研AI框架尚未成为主流。

更为严重的是，近年来美国持续打压中国高科技企业，把我国大量企业列入实体清单。美国实体清单重点打击对象是超算和人工智能企业，导致这些领域的供应安全受到严重破坏。实体清单不仅限制了涉及美国技术的AI芯片（如GPU、NPU、FPGA等）和硬件的供应，也涉及到AI软件的管制,包括芯片硬件使能层（如英伟达的CUDA/cuDNN、开源算子开发软件TVM、英特尔的OpenCV、Free软件类视觉编解码库FFMPEG和LibJPEG等）、AI框架层（如TensorFlow、Pytorch、Caffe、RAY、TensorRT等）、应用使能层（模型ResNet50、Bert等，开放数据集ImageNet、COCO等）。

时间	被列入实体清单的企业
2015年4月	国家超算广州、天津、长沙等中心、国防科技大学
2018年8月	44家中国企业（8个实体加36个附属机构，航天、中电科系）
2019年5月	**华为**68个相关实体
2019年6月	江南计算技术研究所、中科曙光、海光系
2019年10月	新疆公安厅、19个附属单位，8个商业实体（**大华、海康、科大讯飞、旷视、商汤、依图、美亚柏科、颐信科技**等）
2020年5月	33家中企（奇虎360、**云从科技**、哈工大……）

表2 实体清单（标粗为AI企业）

因此，要清醒地看到，我国人工智能芯片、框架整体发展水平与发达国家相比仍存在较大差距，核心技术受制于人，科研机构和企业尚未形成具有国际影响力的生态圈和产业链，缺乏系统的超前研发布局。

第四章

加快发展我国人工智能计算中心的建议

当前，人工智能计算中心是新事物，全球都在积极探索人工智能算力的大规模建设和有效使用，随着日益增长的算力需求与计算供给能力之间的差距不断加大，我国人工智能计算中心建设的紧迫性已经凸显，并且存在一系列的问题和挑战：

1）我国人工智能计算中心尚未形成国家统一布局，需借助政府力量集中牵引和统一规划；

2）全球地缘政治环境变化深刻影响产业链供应链体系，我国在关键芯片、核心部件及软件等方面的核心技术能力不强，难以自主可控，需加大对人工智能相关的产业链供应链安全的重视；

3）人工智能技术创新及产业变革也带来生态完整性的问题，不同领域、产业链各环节的企业需不断深化垂直整合，亟需加强基础软硬件的协同适配，形成合力；

4）我国面临人工智能顶尖人才、复合型人才缺乏，产学研用协同不足的问题，可通过以产促研、研用结合等手段快速提升竞争力，并对人工智能计算中心的落地及应用发展形成支撑。

因此，应加快推进人工智能计算中心的统筹规划建设。特别是在国家新一代人工智能创新发展试验区建设过程中，各地应充分结合试验区建设要求，本着高质量和集约化建设的原则，积极筹划和指导人工智能计算中心的筹建，提供高性价比的公共算力服务平台，加快形成支撑新一代人工智能科学研究、技术开发、产品研制、应用推广的系统化基础设施体系，推动当地人工智能产业持续、良性、健康的发展，为中国人工智能产业和数字经济的发展奠定坚实的基础。

4.1 强化资源统筹，有序推动重点城市人工智能计算中心建设

人工智能计算中心的建设是一项系统工程，应充分发挥我国集中力量办大事的制度优势，按照国家“新基建”总体要求，加强规划建设、资金保障、资源调配和政策支持等多方面统筹，面向重点城市人工智能创新发展的实际算力需求，规划和建设集中化的人工智能计算中心，用以承载落地大规模AI算力集群。

1）与城市人工智能产业布局协同。考虑城市产业功能区的整体布局，将人工智能计算中心的规划和城市人工智能功能区的规划相结合，尽量在人工智能聚集的产业功能区集约式规划建设人工智能计算中心，以充分发挥人工智能计算中心对产业的带动作用。

2）系统性设计可持续发展。考虑城市人工智能可持续发展的需求，对人工智能计算中心的规模、能效、成本做综合考虑，整体的建

设规模应在考虑当地算力需求的基础上适度超前，建设规模要留出足够的扩充空间，以满足人工智能飞速增长的算力需求，在能效上可考虑引入业界先进的液冷、数据中心智能功耗控制等先进技术，以保证人工智能计算中心在PUE上能处于较为领先的水平。

3）全栈式一体化快速建设。规划应考虑到人工智能计算中心建设落地的时间紧迫性，加快建设将及早对当地经济社会的智能化发展起到加速作用。可考虑利用现有的数据中心和超算中心的计算基础设施叠加建设人工智能算力的落地，如果是完全新建人工智能计算中心，可考虑采用业界模块化一体化数据中心的建设方式，不建议分层解耦，以缩短整体建设周期，使得计算中心能迅速使能产业。

4）配套出台相关支持政策。研究制定算力服务相关激励政策，围绕人工智能计算中心打造公共算力服务平台，鼓励人工智能企业、研究院所、高校广泛应用人工智能计算中心的相关算力，避免人工智能算力的重复建设和分散建设。依托人工智能计算中心构建区域性人工智能生态创新中心等生态孵化组织，通过举办企业赋能、联合创新、技术交流、展会、培训和人才培养等一系列活动，充分引导和释放人工智能计算中心引流算力需求。

4.2 坚持自主可控，合理选择人工智能计算中心技术路线

人工智能计算中心的方案技术选择应综合评估自主可控能力、核心技术指标、软件使能能力、生态建设能力等多种素，充分考虑技术路线的先进性和自主性，避免人工智能发展过程中技术“卡脖子”的风险，合理选择适合城市的人工智能发展的技术路线。

1）确定核心人工智能芯片和产品路线。选择具备自主可控能力的人工智能核心芯片和核心产品作为整体方案选型的基础，选拔具备芯片自主研发能力、自主知识产权的厂商作为核心供应商，服务器CPU部分的选型重点参考中国信息安全评测中心的CPU选型厂家。

2）采用领先技术保证系统先进性。选择人工智能计算集群的核心技术指标和核心性能指标领先的厂商，采用业界领先的人工智能算力基准测试程序进行评测，保证在主流的人工智能场景下具备领先于业界的算力性能，以保证人工智能计算中心的技术先进性。

3）软硬协同发展自主可控人工智能软件。选择具备自主可控的人工智能软件平台的厂商，尤其是在人工智能系统软件、人工智能开发框架以及人工智能使能平台方面需要具备自主研发能力和自主知识产权，确保人工智能的技术发展牢牢掌握在自己手中。

4）加强应用使能降低使用门槛。选择具备优秀应用使能能力和具备良好生态的厂商，从应用平台的开放性、易用性和应用生态的构建情况进行综合评定，选择适合当地城市产业应用发展的技术平台，以促进城市人工智能应用生态的发展。

5）重视全栈方案设计建设能力。选择具备大规模集群建设经验的厂商，优先考虑具备软硬件全栈建设能力和经验的头部厂商，人工智能集群建设方案复杂，技术难度高，选择具备丰富建设经验的厂商，有效避免方案建设过程中的技术风险。

6）建设云平台统一管理和服务。选择具备大型公共云平台建设与运营经验的厂商，从平台服务方案完备性、技术先进性、市场占有率和运营运维能力进行综合考量，以确保云平台为人工智能应用提供足够稳定优异的云底座。需要强调的是，平台还需具备最新一代技术的持续演进能力，如边缘计算、容器、大数据、量子计算等，这些通用技术将与人工智能技术持续融合，使能企业愈加丰富的AI+融合应用场景。

4.3 加强数据开放共享，促进人工智能计算中心赋能场景应用

人工智能计算中心的长期发展，离不开人工智能场景应用所需的各类数据资源，应加快构建科学合理的数据开放共享机制，降低人工智能训练获取数据的门槛，从而进一步带动人工智能计算中心的算力使用。

1）建设数据安全开放机制。以政府部门

为重点，大力推动数据开放、安全共享机制建设和实施，推进国家就业、社保、地理、环境、生态、交通数据的开放共享，支撑人工智能与政府服务的融合，提升政务服务水平。

2）整合数据资源并开放共享。稳步推进教育、医疗、能源、公共安全等领域数据的内部整合、共享与对外开放，制定数据资源清单和开放计划，支持相关企事业单位联合人工智能企业围绕应用场景开展人工智能服务，鼓励优质机构人工智能服务能力和资源向地方开放。

3）建立数据开放运营机制。通过公共数据的公开共享，引导企业、行业协会、科研机构、社会组织等主动采集并开放数据。构建安全有序的数据交易环境，推动地方政府建立数据交易平台，确保数据安全，规范交易流程，把控交易数据质量。

4）完善数据的流通与协同。配合新型信息基础设施实现数据的流通与协同，规范不同行业之间数据协同的对接标准，推动我国政府数据开放生态体系的建设与发展，建立数据确权机制，健全数据安全和隐私保护法规，加速实现社会高效治理和高效运转。

4.4 重视人才培养，依托人工智能计算中心打造区域创新人才高地

人工智能计算中心可以为区域的科研创新和人才培养提供丰富的算力支撑，同时人工智能计算中心的长期持续发展也依靠区域科研创新的不断发展和区域人才的持续培养。应在建设人工智能计算中心的基础上，配合一系列的科研创新和人才培养的激励政策，着重打造城市科研创新和人才培养平台，形成区域人工智能人才发展的有利条件，进而形成城市人工智能发展的良性循环。

1）推进建立新型科研组织。鼓励高校院所联合行业龙头企业，采用政产学研合作模式创建一批人工智能重点实验室、研究院等创新科研组织，围绕人工智能计算中心的充沛的算力资源开展人工智能技术研发、科技成果转化等工作。

2）加强人工智能学科建设。支持省内高校设立人工智能学院或研究院，引导龙头企业、科研院所等参与高校的人工智能学科建设，增强人工智能基础理论与前沿技术领域研究力量，推进人工智能与医学、农学等学科的交叉融合。

3）制定人工智能引才计划。大力引进人工智能基础理论、关键技术等领域的高端紧缺人才和高水平创新团队，加快引进人工智能领域的青年创新型人才。依托重大科技专项、博士后科研流动站、博士后科研工作站、博士后创新实践基地、博士工作站等重大人才平台和基地，在人工智能重点发展领域培育一批具有

发展潜力的人工智能青年领军人才与科学家。

4）壮大高端复合性人才队伍。鼓励人工智能企业通过长短期聘用、项目合作、技术咨询等柔性引才方式，灵活引进高端人才。针对人工智能领域高端紧缺人才，开辟人才绿色通道，在人才落户、子女教育等方面给予倾斜支持。推动企业加强人才自主培养，形成一批掌握人工智能应用的复合型人才和团队。

4.5 完善运营机制，实现人工智能计算中心的健康发展

人工智能计算中心的建设和运营同样关键，如果缺乏一个强有力的人工智能计算中心运营主体和运营机制，人工智能计算中心的算力可能会出现闲置，也无法发挥人工智能计算中心对于人工智能产业发展的带动作用，尤其对于政府投资和建设的人工智能计算中心，运营方的选择尤为关键，其中运营公司的能力矩阵建议如下：

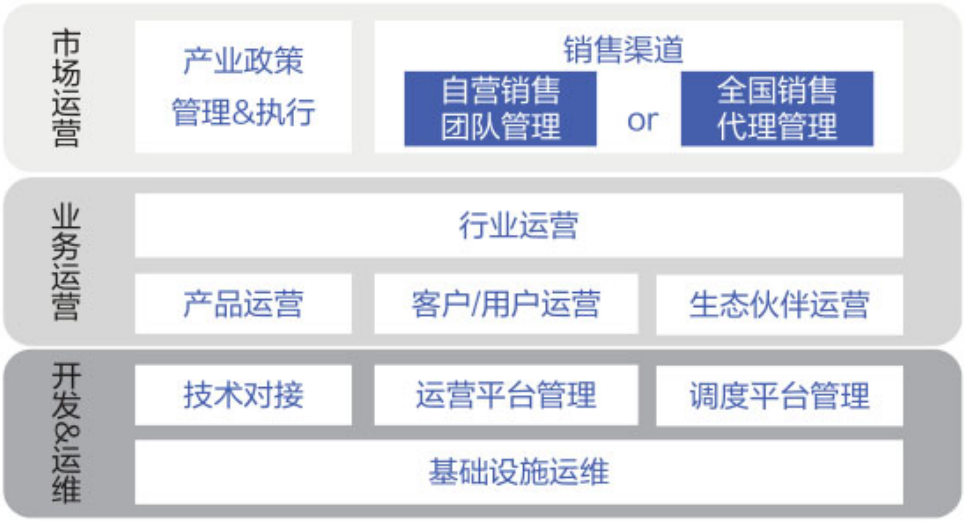


图7 运营公司能力矩阵

1）市场运营；

- 政府或国资背景，有利于省内产业整合和政策执行落地；
- 销售渠道：具备自建全国销售渠道或全国渠道招募能力，快速实现全国覆盖；

2）业务运营；

- 具备类似行业运营经验，了解算力需求潜在客户群和客户的主要诉求，以及了解业界同类产品商务体系，便于迅速开展业务；
- 具备一定的ISV生态伙伴资源，能够满足最终客户诉求；

3）开发和运维；

- 开发调测能力：具备一定的技术能力，能支撑生态伙伴的对接，以及具备一些微小功能开发的能力，能够开发一些工具简化运营及对接工作；
- 运维：机房、设备等日常维护基础能力；

4.6 强化市场作用，提升人工智能计算中心协同应用水平

城市人工智能中心的建设要赋能产业智能化升级，需要构建开放共享、安全的人工智能开放市场，降低开发者的开发周期和应用门槛，丰富城市人工智能资产类型和数量，并通过开发者和服务商智能匹配、交易形成人工智能自增长的数字集市，打造包含开发者和消费者在内的城市人工智能生态群落，围绕算法、模型、场景模板、API和镜像等多种形式的人工智能资产共享及交易，实现人工智能生产力要素的流通与协作，构筑城市人工智能生产和消费的大循环，从而提升人工智能计算中心的协用应用水平。

1）建立互联网化的人工智能资产流通市场。基于人工智能开发者生态社区，提供先进算法、可信模型和在线API等形式的资产共享、交易功能，为高校科研机构、产业人工智能场景应用开发商、解决方案集成商、企业及个人开发者等数字经济参与者，提供开放共享的交易环境和配套政策，以市场化的手段激发人工智能要素市场活力。

2）完善人工智能资产市场的应用通路，加快人工智能应用进程。为人工智能开放市场的参与者提供身份认证、权限管理、计费策略等基础服务，打通人工智能资产和人工智能计算中心的应用通路，使能人工智能应用高效再训练、端到端部署和应用效果监控预警等全流程多环节闭环，持续提升人工智能在生产系统的泛化能力和推理效果。支持人工智能应用能力的API化输出，构建城市人工智能统一接口，向经济主体提供开放人工智能接口，打造开放的新型人工智能基础设施。

3）加强投入保障人工智能资产市场的安全性。采用可信评估技术监督人工智能生产者的发布环节，从供给侧保障人工智能资产可信。通过人工智能模型安全技术防止人工智能资产被盗取、被投毒，确保人工智能生产和消费的可信、可控和安全，从而吸引更多经济主体参与城市人工智能生产和消费的大循环。

总之，建设人工智能计算中心不能一蹴而就，需要系统谋划和统筹推动，并结合相应的政策扶持，切实以人工智能计算中心培育带动人工智能产业链的完善和发展，提升对传统企业的智能化转型升级支持力度，从而加快推动人工智能与经济社会的深度融合发展。特别是在推动建设人工智能计算中心的过程中，对人工智能计算中心的投入产出效益评估，不应从短期直接经济效益角度出发，更重要的是要关注人工智能计算中心作为智能化基础设施，对经济社会长期发展的宏观带动作用。

作者
AUTHOR

中国科学技术信息研究所

赵志耘　　党委书记

徐　峰　　政策与战略研究中心 副主任

高　芳　　政策与战略研究中心 副主任

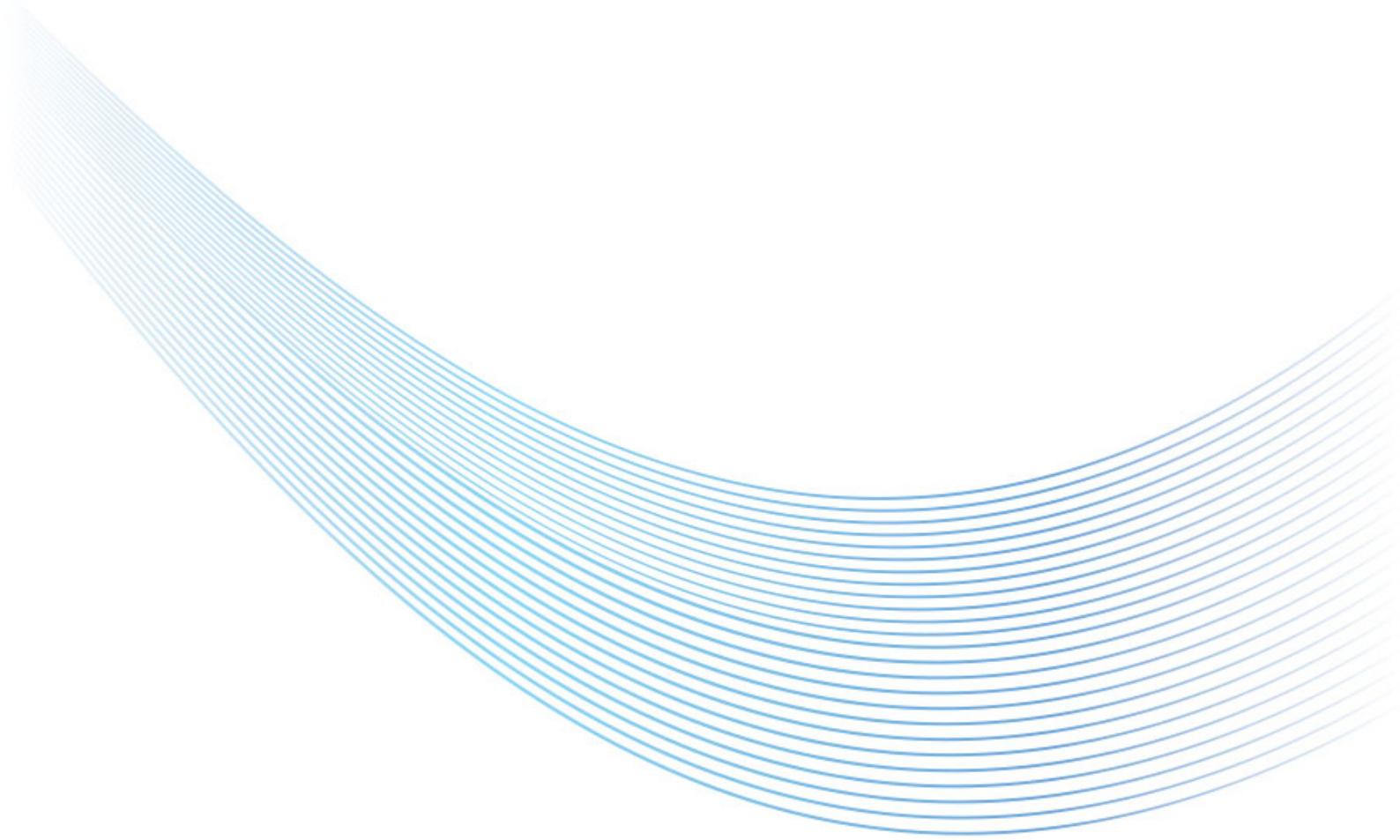
李梦薇　　政策与战略研究中心 助理研究员

中国科学技术信息研究所

中国北京市海淀区复兴路15号，100038

+86 10 5888-2538

www.istic.ac.cn



免责申明