智算中心建设通过领先的体系架构设计，以算力基建化为主体、以算法基建化为引领、以服务智件化为依托，以设施绿色化为支撑，从基建、硬件、软件、算法、服务等全环节开展关键技术落地与应用。

一、体系架构

（一）总体架构

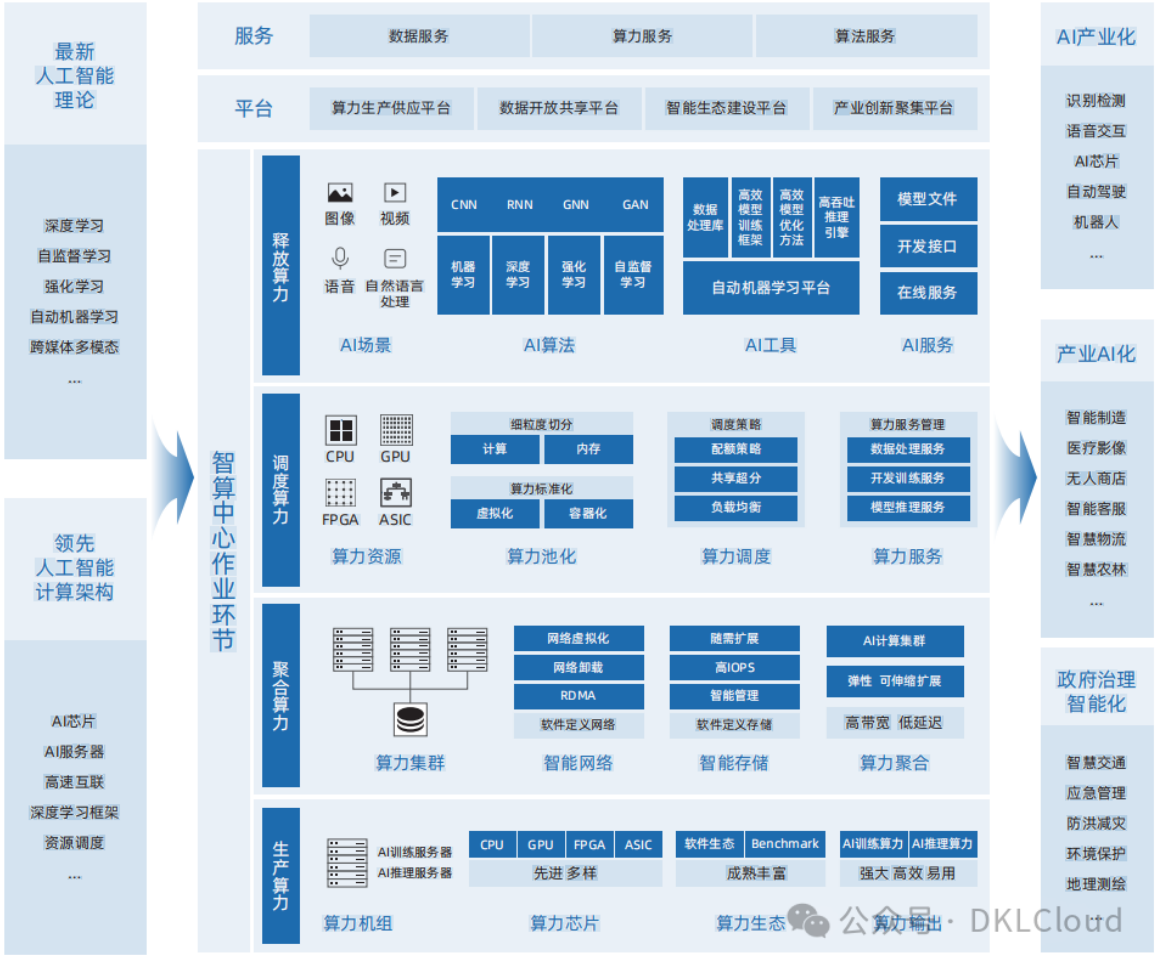

图8 智算中心总体架构

智能算力中心建设白皮书，重点围绕基础、支撑、功能和目标四大部分，创新性地提出了智算中心总体架构。

其中，基础部分是支撑智算中心建设与应用的先进人工智能理论和计算架构；支撑部分围绕智算中心算力生产、聚合、调度、释放的作业逻辑展开；功能部分提供算力生产供应、数据开放共享、智能生态建设和产业创新聚集四大平台，以及数据、算力和算法三大服务；整体目标是促进AI产业化、产业AI化及政府治理智能化。

（二）技术演进

智算中心的发展基于最新人工智能理论和领先的人工智能计算架构，算力技术与算法模型是其中的关键核心技术，算力技术以AI芯片、AI服务器、AI集群为载体，而当前的算法模型发展趋势以AI大模型为代表。

在此基础上，通过智算中心操作系统作为智算中心的“神经中枢”对算力资源池进行高效管理和智能调度，使智算中心更好地对外提供算力、数据和算法等服务，支撑各类智慧应用场景落地。而软件生态则是智算中心“好用、用好”的关键支撑。

1．AI芯片

基于AI芯片的加速计算是当前AI计算的主流模式。AI芯片通过和AI算法的协同设计来满足AI计算对算力的超高需求。当前主流的AI加速计算主要是采用CPU系统搭载GPU、FPGA、ASIC等异构加速芯片。

AI计算加速芯片发端于GPU芯片，GPU芯片中原本为图形计算设计的大量算术逻辑单元（ALU）可对以张量计算为主的深度学习计算提供很好的加速效果。随着GPU芯片在AI计算加速中的应用逐步深入，GPU芯片本身也根据AI的计算特点，进行了针对性的创新设计，如张量计算单元、TF32/BF16数值精度、Transformer引擎（Transformer Engine）等。

近年来，国产AI加速芯片厂商持续发力，在该领域取得了快速进展，相关产品陆续发布，覆盖了AI推理和AI训练需求，其中既有基于通用GPU架构的芯片，也有基于ASIC架构的芯片，另外也出现了类脑架构芯片，总体上呈现出多元化的发展趋势。但是，当前国产AI芯片在产品性能和软件生态等方面与国际领先水平还存在差距，亟待进一步完善加强。总体而言，国产AI芯片正在努力从"可用"走向"好用"。

2．AI服务器

AI服务器是智算中心的算力机组。当前AI服务器主要采用CPU+AI加速芯片的异构架构，通过集成多颗AI加速芯片实现超高计算性能。

为满足各领域场景和复杂的AI模型的计算需求，AI服务器对计算芯片间互联、扩展性有极高要求。AI服务器内基于特定协议进行多加速器间高速互联通信已成为高端AI训练服务器的标准架构。

目前业界以NVLink和OAM两种高速互联架构为主，**其中NVLink是NVIDIA开发并推出的一种私有通信协议，其采用点对点结构、串列传输，可以达到数百GB/s的P2P互联带宽**，极大地提升了模型并行训练的效率和性能。

OAM是国际开放计算组织OCP定义的一种开放的、用于跨AI加速器间的高速通信互联协议，卡间互联聚合带宽可高达896GB/s。

浪潮信息基于开放OAM架构研发的AI服务器NF5498，率先完成与国际和国内多家AI芯片产品的开发适配，并已在多个智算中心实现大规模落地部署。

3．AI集群

大模型参数量和训练数据复杂性快速增长，对智算系统提出大规模算力扩展需求。通过充分考虑大模型分布式训练对于计算、网络和存储的需求特点，可以设计构建高性能可扩展、高速互联、存算平衡的AI集群来满足尖端的AI计算需求。

AI集群采用模块化方法构建，可以实现大规模的算力扩展。AI集群的基本算力单元是AI服务器。数十台AI服务器可以组成单个POD计算模组，POD内部通过多块支持RDMA技术的高速网卡连接。在此基础上以POD计算模组为单位实现横向扩展，规模可多达数千节点以上，从而实现更高性能的AI集群。

AI集群的构建主要采用低延迟、高带宽的网络互连。为了满足大模型训练常用的数据并行、模型并行、流水线并行等混合并行策略的通信需求，需要为芯片间和节点间提供低延迟、高带宽的互联。另外，还要针对大模型的并行训练算法通信模式做出相应的组网拓扑上的优化，比如对于深度学习常用的全局梯度归约通信操作，可以使用全局环状网络设计，配置多块高速网卡，实现跨AI服务器节点的AI芯片间RDMA互联，消除混合并行算法的计算瓶颈。

AI集群的构建需要配置面向AI优化的高速存储。通过配置高性能、高扩展、多层级的智能存储，为各种数据访问需求提供优化性能。智能存储具备随需扩展功能，实现高IOPS处理能力，支持RDMA技术，同时实现高聚合带宽。

4．AI大模型

超大规模智能模型，简称大模型，是近年兴起的一种新的人工智能计算范式。和传统AI模型相比，大模型的训练使用了更多的数据，具有更好的泛化性，可以应用到更广泛的下游任务中。按照应用场景划分，AI大模型主要包括语言大模型、视觉大模型和多模态大模型等。

自然语言处理是首个应用大模型的领域，BERT是大模型的早期代表。随着大模型在自然语言的理解和生成领域成功应用，推动了语言大模型向更大的模型参数规模和更大训练数据规模的方向发展。当前，语言大模型的单体模型参数已经达到千亿级别，训练数据集规模也达到了TB级别，训练所需计算资源超过1000PetaFlop/s-day（PD）。业界典型的自然语言大模型有GPT-4、源、悟道和文心等。自然语言大模型已经广泛应用于个人知识管理、舆情检测、商业报告生成、金融反欺诈、智能客服、虚拟数字人等场景，同时也出现了一系列的创新应用场景，如剧本杀、反网络诈骗、公文写作等。

在语言大模型大获成功之后，相关技术和方法也被引入计算机视觉领域，通过构建更大的预训练模型，使其可以适用于目标检测、语义分割、异常检测等广泛的视觉任务。

在算法架构上，视觉大模型采用以Transformer架构为主体的神经网络架构和自监督的训练方法以及十亿级的无标注图片数据进行训练。当前业界已经出现了越来越多的通用视觉大模型和面向特定领域的视觉大模型。视觉大模型也已广泛应用于自动驾驶、智能安防、医学影像等领域。

随着大模型技术在语言、视觉等多个领域的应用，融合多个模态的多模态大模型也逐渐成为了业界关注的重点。基于多模态大模型的以文生图，文生视频技术也迅速发展，代表性模型有DALLE-2、Stable Diffusion 3 和Sora等。由于多模态大模型的快速发展，AI内容生成（AI Generated Content，AIGC）已成为下一个AI发展的重点领域。

5．智算OS

智算OS，即智算中心操作系统，是以智算服务为对象，对智算中心基础设施资源池进行高效管理和智能调度的产品方案，可以使智算中心更好地对外提供算力、数据、算法、智件等服务，有效降低算力使用门槛，提升资源调度效率，支撑各类智慧应用场景落地，是智算中心的“中枢神经”。

智算OS主要由三层架构构成，分别为基础设施层、平台服务层、业务系统层。基础设施层主要实现将异构算力、数据存储、框架模型等转化为有效的算力与服务资源，算力资源池能够聚合并进行标准化和细粒度切分，以满足上层不同类型智能应用对算力的多元化需求，并通过异构资源管理和调度技术，提升可同时支撑的智算业务规模。

平台服务层主要提供AI训练与推理服务、数据治理服务、运营运维服务等，并通过智算OS实现自动化、智能化，有效摆脱人力束缚，促进算力高效释放并转化为生产力。业务系统层是面向用户端的统一服务入口，向下整合各层级核心功能，为用户提供多元化、高质量的智算服务，满足生产中不同阶段、不同场景的智算需求。

智算OS以智算中心为载体，通过建设多元、开放的智算平台，融合国际、国内先进人工智能技术，形成标准化、模块化的模型、中间件及应用软件，以开放接口、模型库、算法包等方式向用户提供如行业大模型、自动驾驶、元宇宙、智慧科研等人工智能服务，促进人工智能技术成果的开放与共享，构建开放的智算生态。

6．软件生态

基于业界主流、开源、开放的软件生态建设智算中心，是智算中心能够满足前沿AI计算需求、提升AI创新和生产效率、丰富行业AI应用、促进AI产业快速发展的主要前提。深度学习的加速计算始于GPU，构建于GPU之上的CUDA软件栈为深度学习的算法开发提供了极大的便利。**CUDA软件栈为深度学习的应用开发和计算加速提供了丰富的底层支撑，如张量和卷积计算加速、芯片互联通信加速、数据预处理加速、模型低精度推理加速等。**在此基础上，学术界和工业界已经构建庞大的开源、开放、共享的AI软件生态，有力促进和加速全球AI技术与应用的蓬勃发展。

深度学习框架是当前主要的人工智能算法开发工具。其中TensorFlow和PyTorch的使用较为广泛。TensorFlow因其丰富的模型开发和应用部署组件而在工业界广泛应用，PyTorch则由于其易用性和灵活性在前沿算法开发和学术创新研究领域取得了领先地位。国内的AI科技公司也在开发和推广深度学习框架。其中百度开发的飞桨提供了兼具灵活和效率的开发机制，并联合开源社区打造了一系列覆盖主流产业应用需求的工业级模型，目前在国内已得到较多的采用。

在深度学习框架之上，为了适应计算机视觉任务、自然语言大模型等特定场景的应用开发需求，业界构建了一系列的开源开发库，比如面向目标检测任务的mmdetection、面向大模型训练任务的Megatron-LM、DeepSpeed，以及面向自监督学习的VISSL等。

这些软件库进一步简化了模型训练和应用开发的难度，已成为当前人工智能计算的重要软件底座。业界前沿的知名AI算法，如ChatGPT、DALLE-2、StableDiffusion等都是在这样的架构下实现的。随着国产AI计算产业的快速发展，各厂商也高度重视并投入软件生态建设，力求实现好用、易用的软件开发和应用生态。但总的来说，当前国产AI计算软件生态起步较晚、基础薄弱，还要持续不断加大投入，在各个层面加强建设完善。

（三）建设架构

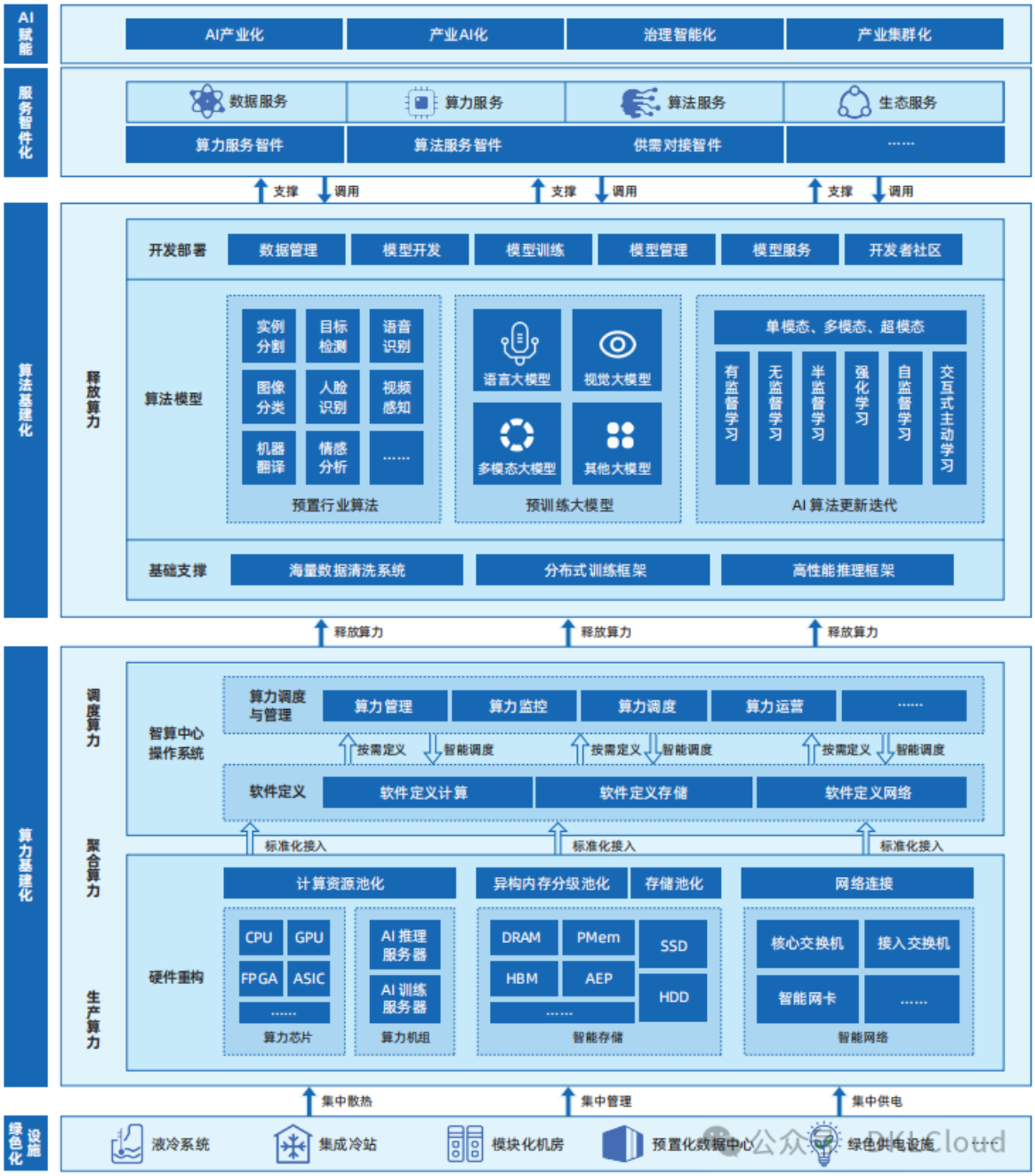

图9 智算中心建设架构

在智算中心总体架构的基础上，聚焦智算中心建设与应用中涉及的关键技术，进一步提出智算中心建设架构。**智算中心建设架构由四大关键环节组成，分别是算力基建化、算法基建化、服务智件化、设施绿色化，“四化”相互支撑、相互协调，共同构建起智算中心高效运行体系。**

同时，在总体架构三项服务、三项目标的基础上，进一步拓展丰富智算中心的功能和目标，实现对外提供数据服务、算力服务、算法服务、生态服务四大服务，支撑达成AI产业化、产业AI化、治理智能化、产业集群化四大目标。

二、技术路线

智算中心建设的关键技术涉及与其建设和应用相关的各类基建、硬件、软件，体现在智算中心算力基建化、算法基建化、服务智件化、设施绿色化过程中。

（一）以算力基建化为主体

以智算中心为代表的算力基础设施能够有效促进AI产业化和产业AI化，是支撑数字经济发展的重要基础底座。为了让AI真正地赋能到千行百业，并推动产业数字化转型发展，智算中心要具备对外提供高性价比、普惠、安全算力资源的能力，使AI算力像水、电一样成为城市的公共基础资源，供政府、企业、公众自主取用。算力基建化供给成为支撑产业转型升级以及创新发展的刚性需求和必然选择。

1．面向潜在算力需求，适度超前规模化部署算力资源数据量的爆炸式增长以及万亿参数大模型的出现，使智能算力需求呈现高速增长态势，并为算力基础设施带来巨大挑战。在数据量方面，IDC发布的《数据时代2025》预测，到2025年，全球数据量将达到175 ZB，而中国数据量的平均增速快于全球3%，预计到2025年将增至48.6 ZB，占全球数据圈的27.8%。在模型方面，当前1万亿参数的单体模型需要1EFLOPS级算力（FP16）计算约50天，10万亿参数的单体模型需要10 EFLOPS级算力（FP16）计算约50天。因此在智算中心的规划建设中，需要聚焦当前算力应用需求，同时面向未来数据量和大模型大参数量增长空间，适度超前，部署满足AI训练、AI推理等大规模计算需求的强大AI算力机组，构建算力集群，提供大规模弹性算力。

2．聚焦异构加速技术，提升高性能人工智能计算能力自2012年以来，人工智能训练任务所需求的算力每3.43个月就会翻倍，大大突破了传统以每18个月为周期实现芯片性能翻番的摩尔定律，这对人工智能计算架构的性能提出了更高的要求。AI芯片是生产算力环节的关键组件，为AI训练和AI推理输出强大、高效、易用的计算力。目前，AI芯片主要包括GPU、FPGA、ASIC、类脑芯片四大类，其中类脑芯片仍在探索阶段，因此多元异构芯片成为提升算力的关键手段。主流的人工智能计算架构是以CPU+AI芯片为主体的异构架构，通过将CPU与多种计算单元（如GPU、FPGA、ASIC等）集成，充分融合了CPU等传统的通用计算单元和高性能专用计算单元的优点，可以同时兼顾AI模型的高效训练和精准推理能力。异构架构具有高性能、高效率、低功耗等显著优点，使AI芯片在未来人工智能算法不断迭代更新的情况下，依旧能保持较好的兼容性和可扩展性，在一定程度上延长了AI芯片的生命周期。

3．兼顾软硬一体协同，构建智算中心多元融合型架构人工智能计算场景和计算架构的多元化要求智算中心从硬件、软件、软硬协同等层面开展优化，提供弹性、可伸缩扩展的算力聚合能力，依据不同类型智能应用对算力的不同需求，提供更高效、更便捷的算力调度能力。采用融合架构进行整体设计是智算中心的发展方向。具体而言，在硬件层面，通过硬件重构实现资源池化，结合新型超高速内外部互连技术、池化融合、异构存储介质等，推动多元异构智能算力设施的高速互联，形成高效池化的智算中心，实现多元计算资源高效协同；在软件层面，通过软件定义，将不同的资源池组成专业的服务器、存储、网络系统，实现重构硬件资源池的高效化、智能化管理，使智算中心的业务资源调度更为灵活、运维管理能力更强。在安全方面，智算中心可以依托隐私安全计算等技术，提供完善的隐私和数据保护解决方案，实现计算、存储、网络等多层级、全方位的资源隔离与安全防护。

（二）以算法基建化为引领

建设适度超前的算力基础设施，不仅体现在算力层面，也体现在算法层面，这是释放算力环节的关键。人工智能算法正面临着丰富化、专业化和巨量化的挑战，智算中心通过提供预置行业算法、构建预训练大模型、推进算法模型持续升级、提供专业化数据和算法服务，让更多的用户享受普适普惠的智能计算服务。

1．面向千行百业发展需求，提供多类型预置行业算法AI落地面临开发成本、技术门槛高的难题，算法模型平均构建时间为3个月，同时算法还需要快速的迭代，再加上AI新算法、新理论层出不穷，行业用户的智慧转型存在着巨大的技术壁垒。智算中心应围绕政务服务、智慧城市、智能制造、自动驾驶、语言智能等重点领域，在AI平台内预置实例分割、目标检测、边缘检测、图像分类、人脸识别、视频感知、自动问答、机器翻译、舆情分析、情感分析、语音识别、协同过滤、交通路线规划等常用行业算法模型，并从硬、软件对行业算法做性能优化，从而帮助各行各业智慧应用加速落地，推动行业智能化转型加速。

2．面向模型即服务应用需求，构建大规模预训练AI模型在产业AI化和数实融合的背景下，当前的行业做法是针对每一个场景都做一个模型，即“有1万个场景就有1万个模型”。然而随着以BERT、GPT-3、DALL·E、源1.0等为代表的高泛化能力和高通用性的大模型的出现，一个模型可以覆盖众多场景。“预训练大模型+下游任务微调”的AI工程化模式已成为业内共识，层数、隐向量长度、前馈网络尺寸持续增长，参数规模迅速从亿级增长到百万亿级。

在充足数据和算力的支持下，大模型可以充分学习文本、图像等数据中的特征。智算中心应通过部署大模型所需要的训练、推理和数据处理系统，构建出不同功能、不同模态的大模型（如自然语言处理大模型、视觉大模型、多模态大模型等），从而更加快速地生产出专业的技能模型，并在更多专业场景中实现小型化、轻量化的落地运作。

3．面向可持续化发展需求，推进AI模型不断演进升级从感知机到深度神经网络，从全连接网络到模型剪枝、知识蒸馏、注意力机制，从有监督学习、无监督学习到强化学习、自监督学习，人工智能理论算法模型在持续深化发展中。当前，人工智能算法正从单模态、有监督学习向多模态、自监督学习演进。自监督学习无需标注数据，可以直接从无标签数据中自行学习，极大降低了人工标注成本。

多模态学习更贴近人类对多感知模态的认知过程，通过学习多种模态的数据，可以突破自然语言处理和计算机视觉的界限，在图文生成、看图问答等视觉语言任务上具有更强表现。随着人工智能相关技术和应用需求的不断升级，智算中心所提供的算法模型也应持续迭代升级，与时俱进，保持算法模型的先进性。未来，人工智能算法将朝着多模态、交互式主动学习、规划、实践的方向发展，以期实现真正的认知智能。

4．面向算法高效调用需求，提供专业化开发部署支撑智算中心除了提供深度学习、强化学习等常见AI算法模型外，还应提供专业化基础支撑和开发部署服务能力，以支撑AI算法模型的便捷调用和部署。为了满足算法模型对大规模高质量海量数据集的需求，智算中心应搭载海量数据清洗系统，提供全流程自动化数据处理系统，实现智能高效的数据处理和过滤。为了满足AI算法模型高效训练和使用的需求，智算中心在基础支撑层面应部署分布式训练框架、高性能推理框架，在开发部署层面应提供数据管理、模型开发、模型训练、模型管理等关键模块，以模型API服务、领域模型、工具包、会话式开放框架、开发者社区等形式，形成强大的AI算法服务支撑能力。

（三）以服务智件化为依托

随着人工智能应用场景持续拓展和开发用户不断普及，对智能计算需求大幅提升、算法模型功能不断强化的同时，人工智能算法开发和模型训练正在从专业化、高门槛向泛在化、易用型转变，智算中心的发展将由传统的硬件、软件向“智件”升级拓展。“智件”是指智算中心提供人工智能推广应用的中间件产品和服务。

传统用户进行人工智能应用时，除了需要提供业务数据，还需提供算法模型并进行代码开发，“智件”的构建可以改变这种服务模式，通过可视化操作界面，以及低代码开发甚至无代码开发的模式，为用户提供功能丰富、使用便捷的人工智能算力调度、算法供给和个性化开发服务，实现“带着数据来、拿着成果走”的效果。

1．提供多元算力调度服务，实现算力调度“智件化”算力是智算中心提供的核心产品和服务。面向不同用户的不同算力需求，智算中心应提供“智件化”算力服务，让用户无需关注底层算力芯片和技术细节，通过用户交互界面，选择业务场景类别、算法模型大小等参数，获得不同算力需求下的计算时间预估、服务费用测算等针对性算力服务方案。

一方面，算力服务虚拟化，弱化底层算力芯片供给的技术差异性，为用户提供标准化的算力供给服务。

通过抽象芯片架构并融合算力特性将提供底层计算能力的GPU、FPGA、ASIC等AI芯片进行统一管理和调度，以PFLOPS、EFLOPS作为计算能力单位向用户提供算力服务，让用户可以更便捷地调度算力，进行AI应用部署。

另一方面，算力服务协同调度，要强化对外的算力调度与服务能力。在构建全国一体化大数据中心协同创新体系和“东数西算”工程的背景和要求下，智算中心可以作为算力基础单元，通过云服务方式融入全国算力调度体系中，满足更大范围、更强算力调度需求。

2．提供简便算法模型服务，实现算法供给“智件化”人工智能是一门极其复杂的学科，要求应用开发者不仅要有扎实的理论功底，还要有高超的编程技术，门槛极高。算法模型是人工智能应用的灵魂，也是智算中心提供服务的主要输出物。从计算智能到感知智能，再到认知智能，人工智能的应用模型越来越复杂，从公共服务到社会治理再到产业发展，人工智能的应用需求越来越广泛，对人工智能模型和算法的要求也越来越高。

为了缓解人工智能模型训练成本高、技术门槛高的问题，智算中心应加强算法供给服务模式的创新，开发可视化操作界面，用户通过API、模块化代码即可获得所需的人工智能应用效果，减轻代码开发压力，使用户无需关注算法和模型本身的复杂技术细节，只需聚焦相应业务领域的业务逻辑和数据就能实现人工智能应用。用户可以基于“智件化”的算法模型进行探索和创新，开发出适用于各种场景的新型智能应用。

3．提供开放生态环境服务，实现供需对接“智件化”人工智能场景日趋丰富，应用需求和技术供给个性化特征明显，为满足部分用户和场景对于人工智能算法优化、系统优化服务的个性化需求，智算中心应构建开放合作生态，加大数据资源供给，聚焦先进的技术并适配典型场景应用。一方面，加大数据供给，数据是人工智能应用的基础，智算中心应打造数据共享平台，推动计算机视觉、自然语言处理、重点行业领域等高质量公开数据集的汇聚，为用户人工智能应用提供增值性数据服务。另一方面，开放发展生态，围绕满足不同用户个性化人工智能应用需求，智算中心应将其计算平台、资源平台和算法平台对外开放，聚集行业内领先企业的力量，及时响应用户个性化需求，提升智算中心技术能力的同时形成新的产业和生产力。

（四）以设施绿色化为支撑

能耗是衡量智算中心发展水平的重要维度之一。“碳达峰、碳中和”目标背景下，国家和地方持续出台政策，进一步规范数据中心的能耗水平和平均电能利用效率（PUE）。为了进一步降低智算中心能耗，设施绿色化是智算中心建设的必然选择。设施绿色化主要包括设备节能化、能源供给绿色化等方面。

1．采用先进节能技术，全面降低智算中心能耗制冷设备和IT设备是智算中心主要的能耗来源。液冷技术采用冷却液和工作流体对发热设备进行冷却，利用高比热容的液体代替空气，提升了制冷效率，降低制冷能耗。液冷技术是智算中心制冷的主要发展趋势。数据中心采用全栈布局液冷，冷板式液冷、热管式液冷、浸没式液冷等先进液冷技术，构建包含一次侧二次侧液冷循环、CDU等的智算中心液冷整体解决方案，可以进一步降低能耗、降低PUE，实现绿色化。液冷智算中心采用余热回收技术，可以为智算中心自身以及邻近区域供暖，进一步提升能源利用效率。此外，智算中心采用高压直流、集中供电等高效供配电系统、能效环境集成检测等高效辅助系统、智能监控运维系统等绿色管理系统可以进一步降低能耗。

2．采用绿色清洁能源，从源头上实现绿色低碳

一方面，智算中心的大部分业务负载，特别是企业负载，在时间上主要集中于白天工作时段，与光伏、风电的主要发电时段匹配性较高，无需过多储能与调峰，使得智算中心在运用光伏、风电等绿色电力方面具有天然优势。采用绿色电力供给的智算中心综合运用线性规划、混合整数规划、启发式算法等多种能耗管理方法，可以在降低碳排放的同时也节约电价成本。智算中心采用优化调度与需求响应控制策略，还可作为需求侧可调载荷参与电力需求侧响应，不仅提升智算中心自身能源利用效率，而且提升新型电力系统需求侧资源优化配置效率。

另一方面，智算中心所在的建筑物、园区空间大，可以充分利用，发展屋顶光伏、园区风电等可再生能源发电设施，优化能源绿色供给格局。应用分布式光伏发电、分布式燃气供能等技术可以提升智算中心园区绿色化水平。小型智算中心还可以利用模块化氢燃料电池、太阳能板房等技术优化能源供给格局。

应用篇

在识别检测、语音交互、智能客服等智能应用在各行业领域得到了广泛使用，以自动驾驶为代表的高算力需求场景从实验环境逐步走向试点应用阶段，而以元宇宙、智慧科研（AI for Science）为代表的新兴场景也逐渐走进大众视野，并带来无限发展可能。

作为支撑人工智能应用的关键基础设施，智算中心汇聚数据、算力、算法等要素，通过生产算力、聚合算力、调度算力、释放算力等关键环节，实现“以数据输入，让智能输出”，助力AI产业化和产业AI化，让智能计算真正惠及经济社会发展。

一、智算中心激发AI产业化创新活力

（一）自动驾驶

自动驾驶是汽车智能化和自动化的高级形态，作为AI技术备受关注的重要落脚点，被公认是汽车出行产业的未来方向之一。自动驾驶场景的实现，需要通过感知融合、虚拟路测（模拟仿真）、高精地图、车路协同等核心技术将数字世界与实体路况进行深度融合，基于人工智能技术，让车辆能够像人类驾驶员一样准确地识别车道、行人、障碍物等驾驶环境中的关键信息，并及时对周围运动单元的潜在轨迹做出预判。

自动驾驶落地需要超大AI算力支持自动驾驶需要通过对车身多个传感器的数据进行感知和融合，并在此基础上对自动驾驶车辆的行为进行决策和控制，其中涉及大量AI算法、机器视觉与传感器数据整合分析、面向各类算力平台及传感器配置方案的适配能力等。

为了提升自动驾驶系统的感知和决策性能，当前通行的做法是在数据中心端基于海量的道路采集数据来进行感知模型训练和仿真测试。随着AI技术的发展，通过AI算法对多传感器的数据以及多模态的数据进行融合感知，已经成为了当前主流的发展趋势。另外自监督大模型的技术也在逐步地引入到自动驾驶场景中。

这都使得自动驾驶感知模型的训练算力消耗远大于一般的计算机视觉感知模型。比如，Tesla构建的L2级别的FSD自动驾驶融合感知模型的训练使用了百万量级的道路采集视频片段，算力投入约为500PD。随着自动驾驶级别从L2到L4的提升，对算力的需求将进一步提高。

算力供给是自动驾驶系统得以大规模落地和进一步商业化的前提条件。自动驾驶产业的集成化、规模化发展需要由智算中心提供超大算力、先进AI算法等支撑。智算中心提供的普惠算力可以极大降低自动驾驶所需算力的成本，同时加速自动驾驶新技术与新产品的研发、测试和应用。

（二）机器人

机器人是人工智能技术多领域应用的重要载体，主要分为工业机器人、服务机器人和特种机器人。作为一种重要的智能硬件，随着计算机视觉、机器学习、智能语音等多种智能算法技术的进步，机器人产业也将实现飞速的发展。

《中国机器人产业发展报告（2022年）》数据显示，2022年中国机器人市场规模约为174亿美元，五年年均增长率达到22%，其中工业机器人和服务机器人市场规模均保持增长，二者呈现出齐头并进、快速发展的态势。

“AI算法+AI算力”支撑机器人从量变到质变机器人与新一代信息技术的融合逐渐深入，机器人的感知、计算、执行能力都得到了大幅提升，处理实际问题的稳定性和可靠性也进一步提高，这背后离不开人工智能技术和强大算力的支撑。机器人需要和环境进行交互感知以及决策控制，和环境的交互感知不仅涉及到视觉、听觉等多个模态，也会涉及到不同模态的感知融合，这都需要AI算法作为底层支撑。为了实现相应的感知和决策算法，一般会在数据中心端构建真实世界数据采集→AI模型构建→孪生世界的决策控制模型训练→真实世界验证测试的闭环，来逐步地提升机器人在真实世界的感知和决策能力。

用于学习和训练的数据越多，算法迭代得越完善，机器人的决策准确度将越高。智算中心的算力服务可以为机器人的大规模模型训练和预测提供强大算力支撑，智算中心的算法服务可以实现机器人智能化应用算法模型的敏捷开发和快速训练上线，为机器人产业的高质量发展提供全方位支撑。

（三）元宇宙

元宇宙是基于数字技术进行创造和连接，与现实空间映射交互形成的虚拟空间，是整合多种新技术而产生的下一代互联网应用和数字形态的新型社会体系。元宇宙在5G、人工智能、物联网、AR/VR、云计算、区块链等技术及产品的支持下，为现实世界构建数字化虚拟平行世界，为用户提供沉浸式交互体验，大幅提升各行业生产效率。

智算中心是支撑元宇宙实现的关键基础设施元宇宙的沉浸式体验离不开扩展现实、人工智能、区块链等元宇宙核心技术的支持，对系统的计算、存储、带宽、功耗等都提出了极高的要求，其所需消耗的算力资源也是巨量的。

元宇宙的协同创建、高精仿真、实时渲染、智能交互等环节都需要大量算力做支撑，想要真正迈入虚拟和现实融合的3D互联网时代，元宇宙对算力的需求将呈指数级增长，这远远超过了通用CPU的发展速度。传统以提升CPU时钟频率和内核数量来提高计算性能的方式遇到了瓶颈，形成了巨大的算力缺口。元宇宙从本质上看是对算力的重构，这部分算力缺口需要由智算中心来弥补，从而不断提升元宇宙场景的性能和能效。

1．虚拟数字人

虚拟数字人是可以感知、规划、行动的虚拟形象，由计算机图形学、图形渲染、动作捕捉、深度学习、语音合成等技术生成，具备类人外貌、交互能力等高度拟人化特征，是元宇宙的重要组成部分。虚拟数字人正逐步"闯入"现实虚拟数字人的应用领域非常广泛，按照应用场景或行业的不同，已经出现了虚拟主播、虚拟偶像等娱乐型数字人，虚拟教师等教育型数字人，虚拟客服、虚拟导游等助手型数字人，替身演员、虚拟演员等影视数字人应用。据《虚拟数字人深度产业报告》预测，2030年我国虚拟数字人整体市场规模将达到2,700亿元。

智算中心助力虚拟数字人应对AI算力和算法挑战虚拟数字人相关的建模、驱动、渲染和感知交互均需要巨量的算力支撑。当前，虚拟数字人的建模以基于3D建模软件的手工建模+真人驱动为主。

随着AIGC等AI技术的应用，基于AI算法的自动建模将逐步替代手工建模，成为数字人建模的主要方式。与此同时，基于AI算法的数字人驱动也将逐步替代当前以"中之人"驱动为主的真人驱动方式。与此同时，视觉感知、语音识别和语音合成以及自然语言处理等多种AI算法在数字人中的应用，将推动数字人向"数智人"转变，也是虚拟数字人应用普及的关键。智算中心可以为虚拟数字人制作、感知交互提供强大的算力和算法支撑，加速虚拟数字人产业的商业化落地。

2．数字孪生

数字孪生是指充分利用物理模型、传感器、运行历史等数据，集成多学科、多尺度的仿真过程，以数字化方式创建物理实体的虚拟镜像，通过模拟、验证、预测和控制物理实体全生命周期行为，实现在物理空间的最优决策。数字孪生是构建元宇宙数字空间的基础数字孪生在元宇宙的发展进程中扮演着重要角色，是元宇宙耦合物理世界的基石。

元宇宙的目标是构建一个与现实物理世界高度贴合的甚至是超越现实世界的虚拟世界，因此需要通过海量数据模拟和强大算力来实现1:1的数字空间创造，这个过程中的核心关键就是数字孪生。数字孪生技术能够以极致细节的方式将现实世界映射到虚拟世界中。因此，数字孪生技术的成熟度在一定程度上决定了元宇宙在虚实映射与虚实交互上的发展潜力。

强大算力是数字孪生高效稳定运行的重要支撑数字孪生的应用十分广泛。例如，数字孪生城市可以在虚拟世界模拟仿真城市管理、产业发展、消防应急、环境变化等情况，为现实中关键问题的决策提供技术支撑，提升城市规划和城市治理的效率和精准度。

在元宇宙中，大规模、高度复杂的数字孪生空间的构建，以及现实世界和数字世界的实时交互，需要有强大且物理准确的高精度仿真算力和实时高清3D渲染算力作为支撑。随着AI技术的发展，基于AI算法的高精仿真逐步替代了传统基于数值求解算法的仿真系统，成为了数字孪生系统的核心底层支撑技术。智算中心可以为大规模数字孪生提供专业化的算力和应用支持，支撑数字孪生空间的实时创建、复杂模型的高效运行，以及逼真仿真环境的快速生成。

二、智算中心助力产业AI化走深向实

（一）智慧医疗

国家统计局《2021年国民经济和社会发展统计公报》显示，2021年全年总诊疗人次85.3亿人次，基本医疗保险覆盖13.6亿人。然而，各个地区医疗服务水平参差不齐，医疗服务资源不均等现象普遍存在，基层患者尤其是偏远地区的患者难以获得高质量的医疗救治。

AI辅助诊断助力解决诊疗"三大难题"当前，医疗诊断主要面临三大挑战：

一是数据量巨大。粗略估算诊疗人次所对应的就医环节及相应的医疗数据质量，加上血压、心率、体重、心电图等医疗监测数据，规模早已突破TB级，并且以"秒"为单位持续更新叠加，需要强有力的算力支撑平台。

二是数据结构多元。不仅包含大量医学术语、专业名称，还包括文档、影像、视频等非结构化数据，对AI服务器等新型智能计算硬件要求较高。

三是数据实时处理要求高。医疗服务中存在大量时间性强和决策周期短的应用场景，如临床中的诊疗和用药建议、健康指标预警等，对在线计算、实时处理的需求显著，亟需构建强大的算力平台支撑基于医疗健康领域数据规模化知识图谱。

医疗机构通过引入AI辅助诊疗，可实现诊断、治疗工作的智能化。从算力需求看，人工智能辅助诊疗应用涉及海量图形数据的处理，所需的算力要求较高。智算中心具备的强大算力可以支持大规模、高难度的模型训练，全方位支撑海量医疗影像数据的分析挖掘和精准诊断，能够有效缩短诊断时间，提高诊疗效率。

（二）文娱创作

近年来，AI在文娱创作方面有诸多突破，通过融合人工智能、认知心理学、哲学和艺术等多个学科，可完成诗词、绘画、音乐、影视、小说等创作。

人工智能正在逐渐改变文娱创作的发展范式对艺术家来说，灵感极为可贵并难以捕捉，当文娱创作遇上人工智能，整个行业迸发出了全新的生机和活力。AI技术将是未来数字化创作的重要生产工具。当前出现的创作生态可分为专业生成内容（Professionally Generated Content，简称PGC）、用户生成内容（User Generated Content，简称UGC）、AI辅助生产内容和AIGC。

其中，PGC和UGC都是以人为主体的创作模式，PGC是由专业人士进行内容创作，成本较高且产能有限；UGC降低了生产成本，满足了个性化需求，但存在不可控因素。从长期来看，数字内容生成的需求会愈发强烈，但是人脑处理信息的能力有限，当以人力为主的内容生产潜力逐渐消耗殆尽，以AI为主的内容生产模式将弥补数字世界内容供需的缺口。Gartner数据显示，到2023年将有20%的内容由AI创作生成，预计到2025年生成式AI产生的数据将占所有数据的10%。

AIGC将成为数字内容生产的长期发展方向AIGC是一种通过生成对抗网络、深度学习、大型预训练模型等人工智能技术挖掘数据中的规律，并通过适当的泛化能力生成相关内容的技术。深度学习技术（如深度学习模型CLIP等）的突破为AIGC商业落地提供了可能，而数字内容、数字资产等的快速发展又进一步加速了AIGC的应用与优化。

利用AIGC技术可以生成多种模态的数字作品，如AI写作（文本）、AI绘画（图像）、AI作曲（音频）、AI换脸（视频）等。同时，AIGC技术也可以实现由文字生成图像、文字生成视频、图像/视频生成文字等跨模态创作，以及Game AI等各类综合型场景创作。AIGC的出现使数字内容创作的生产效率和互动性得到了进一步提升。随着人工智能技术的不断升级以及算力、数据、算法等要素的持续迭代，未来AIGC技术将持续赋

能各类文化创意、生产生活，为数字内容生产带来巨大变革。AI大模型和开放平台为文娱创作提供技术支撑随着各类AI大模型及支持开发者创作的各类AI开源平台的陆续上线，用户可以获取涵盖开源模型API、高质量中文数据集、模型训练代码、推理代码、应用代码、面向AI芯片的模型移植开发等内容的多场景服务。

大模型开放平台的出现极大地降低了文娱类AI应用的开发门槛，即使是几乎没有任何编程经验的文娱创作者，通过在平台上进行简单学习，也可以快速实现文娱类AI应用的开发。AI大模型和AI开源平台作为智算中心算法基建化的重要构成，配合其强大的算力资源，将为创作者打造一片创作的乐土。

（三）智慧科研

AI技术成为继计算机之后，科学家新的生产工具，并催生出了新的科研范式AI for Science。科学家们用AI技术去学习科学原理，根据实验或者计算产生的数据对所求解的科学问题进行建模，从而使复杂问题得到有效解决。近年来，AI也被证明能用来做规律发现，帮助人类从大量的复杂数据中，抽取一些人类观察不到的高维信息和高价值规律，不仅在应用科学领域，也能在自然科学领域发挥作用。AI for

Science 不仅带来了科研效率的显著提升，还能降低科研成本，让更多人都能参与到科学研究中来。

1．生命科学

随着大数据和人工智能的发展、普及和成熟，越来越多的科学研究从假设推动的范式向数据驱动的范式转变，利用大数据和计算机技术挖掘科学洞见。在生命科学领域，通过采用深度学习方法处理海量数据，已经在蛋白质结构预测等领域实现了落地应用。

蛋白质作为生命活动的主要承担者，长期以来都是生命科学工作者研究的重点，其中确定蛋白质的三维空间结构尤为重要。受困于计算量庞大、计算准确度有限，蛋白质三维结构预测领域近年来进展较为缓慢。

采用传统的冷冻电镜三维重构方法，实验仪器昂贵，且图像重构需要耗费大量计算力，而采用传统的分子动力学结构预测计算方案，在平均10300的搜索空间枚举蛋白质的可能构型，需要极高的算力和漫长的计算时间，因此在过去50年的时间，仅有17%的人类蛋白质组得到结构解析。

在智能算力的支持下，DeepMind开发了基于注意力机制深度神经网络的AlphaFold2模型，通过对当前已经测序的数十万蛋白质结构数据和数百万蛋白质序列数据进行学习，实现了端到端直接预测蛋白质的三维结构，并取得了突破性进展，预测结果准确率达到了92.4%。相较于使用费用高昂的实验仪器，单个蛋白结构的预测时间缩短到了分钟级。AlphaFold2的开发是以巨量算力为支撑，具体来说，其训练数据准备消耗了约2亿核时的CPU算力，训练过程消耗了约300PD的AI算力。

2．大规模分子模拟

分子动力学模拟通过求解原子运动的经典力学牛顿方程对相空间进行采样，可以研究体系在相空间的演化过程，还可以通过统计方法得到体系在非零温度下的各种性质，是当前材料和生物化学领域最常用的计算研究方法之一。

近年来，借助神经网络从大量数据中获得规律的优势，将第一性原理计算结果作为训练数据，利用神经网络训练构建势函数的方法引起了广泛的关注。该系列方法从上世纪90年代开始，经过二十多年的发展，在准确性、可扩展性等方面得到了提升，比较常用的方法有DeePMD、SchNet、GAP、MTP等。

2020年深度势能（DP）团队因“结合分子建模、机器学习和高性能计算相关方法，将具有从头算精度的分子动力学模拟的极限提升至1亿个原子规模”，斩获了当年的戈登·贝尔奖（Gordon Bell Prize）。原子间机器学习势函数已经应用于许多实际研究中，可以用于模拟复杂的、多元素的晶体、非晶、液晶、界面、缺陷和掺杂等实验体系，计算精度接近从头算，计算速度却可以比从头算快数百到上千倍。

3．数值计算

矩阵乘法是许多计算任务的核心，其中包括神经网络、3D图形和数据压缩等。因此，提高矩阵乘法效率将直接作用于许多应用。几个世纪以来，数学家认为标准矩阵乘法算法是效率最高的算法，但在1969年，德国数学家Volken Strassen通过研究非常小的矩阵（大小为2x2）证明确实存在更好的算法。然而，更大矩阵相乘的高效算法仍属于尚未攻克的难题。

DeepMind的最新研究探讨了现代AI技术——强化学习如何推动新矩阵乘法算法的自动发现。基本思路是将发现矩阵乘法高效算法的问题转换为单人游戏，然后训练一个基于强化学习的智能体 AlphaTensor 来玩这个游戏，通过对 AlphaTensor进行调整，专门用以发现在给定硬件（如 NVIDIA V100 GPU、Google TPU v2）上运行速度快的算法。实验结果发现，这些算法在相同硬件上进行大矩阵相乘的速度比常用算法快了10-20%，表明AlphaTensor在优化任意目标方面具备了不错的灵活性。因此，强化学习成为加速新矩阵乘法算法自动发现的一种新思路。

从算力需求看，蛋白质结构分析、大规模分子模拟、数值计算相关应用主要涉及海量数据并行计算和大规模模拟实验，对算力和存力需求较高，属于计算密集型和数据密集型任务。智算中心所具备的算力服务能力极度契合AI for Science相关场景的算力需求，将成为支撑科研高质量、突破式发展的重要基础设施。

建设篇

从建设用途来看，智算中心除充分考虑其普惠性、开放性和集约性外，核心是以高质量、低成本、高性能的AI算力来支撑产业创新、城市发展中的各项智能服务。智算中心建设以总体规划、政企协同、需求牵引为宗旨，聚焦先进的技术和适配典型场景。同时，以智算中心建设和应用带动人工智能产业集群的汇聚，吸引数字化人才，激发人工智能产业的创新活力，推动人工智能产业和区域经济的可持续发展。

一、建设类型与策略

智算中心建设并非简单做好基建即可，还需结合建设基础、当地或区域产业特色，以差异化算力需求为导向，分类引导施策，优化建设方式，改建并行，发展与数字经济相适应的智算中心。

（一）建设原则

政府引导，需求牵引。以政府侧和市场侧实际需求为牵引，以高标准建设、可持续发展为路径，改造存量与优化增量协同推进，引导龙头企业建设高附加值、产业链带动效应明显的重点项目。

开放多元、培育生态。以开放计算为核心，以多元算力融合为方向，推进智算产业核心关键技术的研发标准化、产业化和应用迭代。加强对智算中心关键软硬件产品的研发支持和大规模应用推广，突破关键核心技术，提升智能算力全产业链自主创新能力。

普适普惠、创新发展。以融合架构计算系统为平台，以数据为资源，以强大的计算力驱动AI模型对数据进行深度加工，使智能算力可以像水电一样,成为社会基本公共服务,面向城市各领域应用提供高品质智算服务。

集约高效、节能降碳。坚持集约化、规模化建设方向，加快节能低碳技术研发应用，提升可再生能源利用率，应用节能新技术，减少碳排放，推进智算中心绿色、高质量发展。

（二）依据建设方式分类建设

1.新建智算中心

（1）建设条件

面向京津冀、长三角、粤港澳大湾区、成渝，以及贵州、内蒙古、甘肃、宁夏等全国一体化算力网络国家枢纽节点和数据中心集群，以及人工智能产业领域应用场景多元和科教资源丰富的优势地区，建设智算中心，以智算中心为牵引推动人工智能领域创新要素集聚，打造人工智能产业生态圈。新建智算中心作为新型公共算力基础设施和赋能平台，应支撑国家和区域内重要需求、科研创新和战略任务落地，为AI大模型训练、自动驾驶、生物工程、智能制造、数字孪生、空间地理等人工智能探索应用提供强大的智能算力服务，通过智能算力服务赋能产业升级，带动区域经济发展。

（2）建设方式与策略

加快梯次布局，打造一批城市级智算中心。对于产业智能化发展需求迫切、人工智能产业集聚的地区，可新建围绕人工智能产业需求设计、为人工智能提供专门服务的智算中心，按照适度超前原则配置优质算力资源，提供兼具公有、专用、弹性计算的服务能力，满足不同应用场景和多类型用户的需求，面向当地企业、科研院所等提供科研创新、人才培养、应用孵化、产业发展等服务，打造“易用”“好用”的智算中心。

强化普惠智能算力高质量供给，降低算力使用门槛，推动智能算力服务与物联网和区块链等技术融合创新，打造具有地方特色服务本地辐射周边的智算中心。加强场景赋能，按需建设专业型智算中心。开展面向性能、价格、效益等多方面的测算，形成应用需求供给和可持续的长效动力机制，加快重点行业的智算中心建设，围绕智能经济、智能社会、科研活动、国家重大活动和重大工程等领域的人工智能创新应用场景，加强供需对接，打造特色场景智算中心，发挥倍增效应，做大做强形成规模化应用，带动人工智能和相关产业发展。

2.已建数据中心升级

（1）建设条件

面向北京、上海、广州以及东部经济发达、人口密度大，对数据要素的产生、存储和处理需求高，但面临地区能耗指标紧张、电力成本高、大规模数据中心开发空间受限等问题的地区，对已建数据中心进行智能化改造，推动传统数据中心向绿色高效、智能集约转型升级。改造升级后的智算中心应优先满足国家及当地政务服务、重大项目及重点实验室的热数据处理和汇聚需求，保障城市基本运行和高效治理需求，保障金融、通信、互联网等战略性行业数据汇集和实时响应计算需求，保障科技赋能和产业创新高性能算力需求。

（2）建设方式与策略

以“以旧换新、增减替代”为原则，对已建存量数据中心进行改造升级，加强AI和传统计算的融合。重点将一些冷数据、静态备份数据为主的存储类数据中心，替换为支撑数字经济、人工智能、区块链、工业互联网等前沿产业发展的智算中心。适度利用关闭及腾退的其他老旧落后的自用型数据中心、存储型数据中心、容灾备份中心资源和空间，升级改造为支撑低时延业务应用，服务智慧城市、车联网等重点应用场景落地。

加快传统数据中心节能低碳技术研发推广，提升资源能源利用效率。智算中心具备高功率密度属性，在制冷方面具有更高的要求。目前大多数AI服务器采用的仍是常规风冷模式，部分超过30kW的数据中心采用液冷模式。随着AI服务器功率密度的提升和使用场景的增多，需要在推动已建老旧小散数据中心向规模化数据中心集群或智能化计算中心转型升级基础上，逐步推广液冷技术的应用，促进全产业链绿色低碳有序发展，助力国民经济各行业整体实现“碳达峰、碳中和”的辐射带动作用。

（三）依据功能定位分类建设

1．产业合作平台

（1）建设条件

面向绝大多数无法承担自建智算中心和独立运营费用的企业，由政府主导，通过统一建设高性能、大规模的智算中心，并以租赁形式为有需求的企业提供算力支撑，省去企业投资建设和运营费用。通过平台开放接口的方式，鼓励行业领军企业将开源的算法、开放的数据资源及运营服务等创新要素输送给IT基础相对薄弱的企业，进一步降低人工智能使用门槛，助力各行业智慧化转型升级。

（2）建设方式与策略

借助ICT基础设施企业物理设施建设优势，通过承建智算中心，搭建产业合作平台，集成最新的人工智能加速芯片和存储介质等，使其成为各新兴计算单元进行大规模融合的重要载体，从需求侧刺激硬件重构和软件定义等融合架构技术创新发展。通过推进平台、框架和算法的协同优化，打通人工智能软硬件产业链，打造人工智能算力技术和产业生态。依托人工智能行业领域企业的专精优势，通过成立合资公司等形式参与智算中心建设和运营，借助智算中心平台扩大自有生态优势。

2．产业园区

（1）建设条件

面向各地方政府以云计算、大数据、智慧城市、虚拟现实、人工智能、区块链等技术应用为核心发展方向的顶层规划布局，围绕利用新一代信息技术对农业、工业、服务业进行全方位、全角度、全链条的数字化改造升级需求，通过合力打造面向未来的智算中心、智算产业促进中心等产业配套载体，构建“产业+配套、平台+生态、数字+赋能”数字产业生态，吸引相关技术企业落户本地，逐步促进产业集群规模化发展，立足本地，辐射带动周边，推动数字经济高质量发展。

（2）建设方式与策略

根据城市规模和产业发展定位的需求，以及经济社会发展等因素，由政府为主导，与企业开展合作，以智算中心项目为依托，建设配套产业园区和人才培养平台等，分类给予针对性的优惠政策，吸引人工智能及其相关领域企业和人才向智算产业园区聚集。针对重点行业的特色应用开展试点示范，形成一批可推广的典型应用创新模式。引导有智算需求的企业积极接入智算中心，使用智算中心服务，加速企业集聚和数据共享。

政府根据智算中心运营的特点进行规划与开发，并在此基础上为园区提供政策支持、税收优惠等，加快应用落地，引领塑造产业生态。

二、建设运营模式

为保证智算中心所释放的经济社会效益最大化，需要选择合理的建设和运营模式，保证智算中心的公共属性，实现长效运营，促进有序布局。

（一）主流建设模式

在全国一体化大数据中心协同创新体系构建背景下，地方政府、产业园区、企业等纷纷将智算中心作为培育人工智能产业生态、提升数字经济能级的有力工具，常见的建设模式包括三种。

1．独立投资建设模式

一是政府独立投资建设。政府对建设项目进行直接投资和管理，建设资金主要来自地方政府财政资金、专项债券发行等，建设完成后智算中心所有权归政府所有。出于促进产业发展、优化产业服务的考虑，不同规模的产业园区日益成为智算中心的投资主体，由园区管委会出资建设智算中心。

二是企业独立投资建设。主要由企业联盟、少数企业联合、单独企业等形式进行投资，旨在服务于特定产业发展和特定场景应用。部分负责投资的企业可以同时作为智算中心的建设方，部分负责投资的企业需要联合专业化建设企业进行施工。该模式虽然由企业出资，但是考虑到智算中心的高投入、对于地方经济发展的高影响等因素，应紧密配合国家“东数西算”工程、全国一体化大数据中心协同创新体系等建设指引。

三是高校或科研机构独立投资建设。主要由高校、科研院所、国家实验室等进行投资，建设一般以智能计算平台为主，服务场景相对单一，建设成本比智算中心小。平台可以向师生、研究人员提供免费的算力支撑，服务于科研教育场景，高校和各类科研机构的科研资源叠加智能算力，为基础研究、前沿科学技术研究提供支持。

2．由第三方出资的建设模式智算中心建设的第三方一般为国有控股企业。该模式下，既实现了政府对项目的建设全过程把控和需求的充分对接，还能有效利用相关国有控股公司已有的科技、人力资本、平台资源、市场等优势。智算中心建成后归第三方公司所有，可以由政府承诺用其他项目进行补贴或者置换。具体细分为两类。一种是由地方政府成立新的国有控股公司，专门负责智算中心的建设投资，另一种由地方政府委托或者授权已有的国有控股公司负责出资。

3．基于特殊项目公司的建设运营（SPV）模式

政府与企业共同出资成立智算中心建设运营项目公司，双方在合作框架协议下按比例出资建设智算中心。政府既可以直接投资参与项目建设，也可以通过国有控股公司、下属事业单位等参与项目建设。项目公司需要由政府授权，按照公司化方式独立运作，负责设计、融资、建造和运营等，向政府、企业提供服务或产品并收取费用。该模式优势在于能够节约政府部门的项目建设成本，实现建设资金筹集，同时启用了专业化建设团队，项目管理方式灵活多样，在项目设计、建设和运营中效率较高。

（二）主流运营服务模式

智算中心出现时间尚短，其运营模式极具探索性，可按照运营方、服务类型、服务内容三方面分析。

1．运营方选择

运营主体指具体负责智算中心投入建设使用后的运营服务机构。与投资主体相比，智算中心运营主体类型应更加多元，运营模式也更为灵活，各类主体通过积极探索差异化个性化运营服务模式保障智算中心高效稳定运行。

一种方式为“投-运”一体化，即由项目投资方出资成立实体运营公司，负责管理算力服务和生态服务。团队成员一般包括运营公司自身管理职能部门，算力服务营销人员、技术支持工程师等算力建设方人员等组成。

一种方式是“投-建”合作模式，即由投资方和承建方共同成立新公司，专职负责算力的运营和对外服务等。该模式下，可以形成投资方和建设方的运营联合，实现运营风险共担，特别是考虑到智算中心后期维护存在一定的技术门槛，在此种方式下，可以保证运营的专业性和高产出。

另一种方式是“建-运”一体化，以承建方主要负责运营。具体由承建方成立运营公司，专职负责算力运营和对外服务。考虑到这种模式下由承建方单独承担运营风险，可以由政府给予运营费用补贴，为了约束运营公司经营行为，可由政府对运营公司进行算力利用率等指标的考核。运营收入收益可以由运营方和政府部门共享。

2．运营服务类型

随着人工智能产业不断壮大，应用场景的持续创新，智算中心逐渐走向市场化，服务对象日益多元。一是综合型。以地方政府建设为主，服务于产业发展、科学研究、公共服务等多元场景。该类型一般由地方政府主导建设，有效发挥了智算中心的公共属性。

二是服务于产业发展。多由产业园区或龙头企业、企业联盟主导建设，主要服务于园区及企业的发展，为人工智能产业向更深更广行业应用发展提供算力保障。

三是服务于科学研究。该类型多由高校、科研院所、国家实验室等承担建设，以投资较低的智能计算平台为主，主要是为高校师生、科研人员的科研工作提供算力、算法等支撑。

3．运营服务内容

提供数据服务。

智算中心作为专门服务于人工智能的数据中心，可以为服务购买方提供多元化的数据服务，例如数据存储、数据清洗、数据分析、数据查询、数据可视化等。该服务属于智算中心的基础性服务。提供算力服务。服务购买方无需关注底层算力芯片和技术细节，只需要把计算过程看作“黑箱”，通过选择业务场景、算法模型等，获取服务方案。政府部门、企业、研究机构可以依托智算中心提供的强大算力，驱动AI模型进行数据深度加工，实现AI应用创新。提供算法服务。人工智能以算法作为灵魂，算法同样是

智算中心的主要服务产品。

随着技术的持续精进和场景的持续拓展，人工智能的算法日趋复杂，面临模型训练成本和技术门槛“双高”的问题。在算法服务模式下，有利于购买服务方专注于自身领域的业务逻辑和数据，依托智算中心提供的语音、图像、自然语言处理、决策等领域的算法能力，创新智慧应用。

提供生态服务。

通过智算中心对外提供算力、数据和算法服务，实现了不同主体的线上汇聚，有利于打造开放、共享的生态，实现多方融合性、深度化合作探索。同时围绕购买服务方的共性需求，智算中心的运营主体和技术团队可以发掘研判行业动态和用户需求，提升智算中心的共性支撑能力，引领探索新的业务场景，构筑新的产业和生产力。