建立全国一体化大数据中心的系统挑战初探

陈汉华 王 雄 华强胜 等
华中科技大学

关键词：一体化大数据中心 资源共享 网络传输 安全保障

简介

2020年12月23日，国家发改委、中央网信办、工信部和国家能源局四部门联合发布了《关于加快构建全国一体化大数据中心协同创新体系的指导意见》，指出加快构建全国一体化大数据中心协同创新体系，是贯彻落实党中央、国务院关于“数据是国家基础战略性资源和重要生产要素”决策部署的重要举措。本文从建设全国一体化大数据中心过程中存在的技术挑战出发，探讨资源共享、网络传输、安全保障和应用支撑四个方面的问题。

资源共享是构建全国一体化大数据中心的核心。跨域分布式数据中心的算力和存储等资源呈现高度异构、分布式部署的状况，给支撑复杂多样的应用带来严重挑战。为此，本文结合资源和应用特征，重点讨论异构算力资源、存储资源高效共享的问题。

高性能数据传输是构建全国一体化大数据中心的基础。近年来，我国网络基础设施显著提升，为实现一体化大数据中心和跨域大数据处理提供了可能。本文重点从高效数据传输、云边端融合近数据处理与传输架构和随流检测智能管理网络三个方面讨论网络传输和管理方面的挑战。

数据和计算的安全是建设一体化大数据中心的重要保障。本文针对跨数据中心的数据安全，讨论数据存储的机密性、面向用户选择性共享的选择性加密和基于属性的加密技术；针对跨数据中心的计算安全，讨论基于硬件的可信执行环境（Trusted Execution Environment，TEE）。成功支撑数字经济产业应用是构建全国一体化大数据中心的内在需要，本文从5G、工业互联网、人工智能产业等方面讨论全国一体化大数据中心可拓展的创新应用场景。

建设全国一体化大数据中心是一个系统化的重大工程，在系统层面面临许多开放性挑战，需加快核心关键支撑技术的突破和供给，加快典型应用部署和用户的融入。

跨域数据中心国外现状

近二十年来，国外IT巨头公司均建立了自己的跨域数据中心。例如，截至目前，谷歌在全球范围内已建有23座数据中心，其中美国14座，欧洲6座，南美洲1座，亚洲2座；微软在全球共有16座数据中心，其中美国6座，欧洲2座，巴西1座，亚洲4座，澳大利亚3座。这些数据中心向200多个国家和地区提供包括计算、存储与数据库、大数据和机器学习以及网络等方面的产品和服务。此外，一些科研教育组织也进行了类似尝试。全球网络创新环境（Global Environment for Networking Innova-

tions，GENI）是美国计算机社区在美国国家科学基金会支持下探索的设施概念。GENI为计算机网络和分布式系统的研究提供了大规模虚拟实验条件，促进了网络科学、安全、服务和应用程序中的创新。此外，2002年开始建设的PlanetLab平台被用于计算机网络和分布式系统研究，截至2010年在全球范围内设立了507个站点和1090个节点，产生了较大影响。

云计算数据中心数据传输的接入网和云计算中心网络随着数据中心的部署呈现蓬勃发展之势，同时，边缘计算作为辅助和补充云计算数据中心计算能力的信息基础设施也日益发展，形成了云边一体化的网络和计算架构。例如，亚马逊将AWS（Amazon Web Services）的功能和服务拓展到AWS Local Zones，推出Wavelength的边缘架构，通过调度靠近端侧的数据中心为用户提供近距离毫秒级的低时延快速服务。又如，微软在其公有云Azure的基础上扩展形成了Azure Edge Zones框架。在数据中心连接网方面，亚马逊搭设了覆盖全球的基础网络。谷歌云针对其全球数据中心铺设了专属广域网，包括网络边缘连接。微软拥有并运营全球最大的主干网络之一，其中全球性的复杂网络体系结构跨越超过165000英里（约265541.76公里）。

综上，国际上现有的跨域数据中心的建设目标主要是服务IT巨头公司业务或科研试验。而我国要建设的全国一体化大数据中心旨在为全国数字经济提供普惠式的共享平台，无论是资源共享的规模、网络传输需求，还是安全保障挑战，都是目前国外跨域数据中心平台无法比拟的。

技术挑战

跨域资源共享

异构计算资源共享

跨域数据中心计算资源异构性明显[1]，既有基础算力、超算算力，又有智能算力。基础算力主要是面向传统云计算等应用需求的CPU服务器，超算算力主要是面向科学计算，基于众核处理器的超级计算机；智能算力包括面向人工智能计算的基于GPU、FPGA、ASIC等人工智能芯片的加速计算平台等。为了发挥一体化大数据中心的算力资源共享优势，首先应考虑如何为上层应用程序合理适配跨域异构算力资源，解决异构计算系统面临的挑战，特别是异构系统结构、指令集和编程模型等给软件开发者带来的困难。其次，应考虑如何构建跨域异构算力资源管理机制，研究细粒度分布式异构资源管理。此外，如何结合我国算力资源的现状和优势实现共享也值得探讨。例如，在智能算力方面，截至2020年，中国在全球占比52%，相比美国所占的19%，优势明显，在当前人工智能成为各国科技和产业竞争焦点的形势下，可开展重点面向人工智能应用的一体化智能大数据中心的建设。

算力资源共享管理，需要结合应用服务质量需求和计算模式特征等。首先，跨域算力资源共享要充分考虑应用的延时敏感性。对于延时敏感型应用（例如工业互联网、远程驾驶、远程医疗、车联网、无人机、增强现实/虚拟现实、智慧家庭、沉浸式视频直播、多人在线对战游戏、云桌面等，见表1），应以本地数据中心的算力和边缘算力为主；而对于延时不敏感的应用（例如视频渲染、后台加工、离线分析、报表分析、日志分析、存储备份等）可考虑利用远程数据中心的算力资源。其次，跨域算力资源共享要结合上层应用涉及的计算模式的耦合特征。对于松耦合式计算（例如搜索合适的设计方案、

表1 典型的延时敏感型应用对网络性能的需求

延时敏感型应用	带宽要求	时延要求
工业互联网	低	500 us~50 ms
远程驾驶	25 Mbps~6 Gbps	5~20 ms
远程医疗	25 Mbps~6 Gbps	5~20 ms
车联网	低	2~20 ms
无人机应用	200 kbps~100 Mbps	50 ms
增强现实/虚拟现实	约10 Gbps	10~20 ms
智慧家庭	100~400 Mbps	20~40 ms
沉浸式视频直播	>200 Mbps	20 ms
多用户在线对战游戏	200~300 Mbps	<50 ms
云桌面	20~100+ Mbps	5~15 ms

理解参数空间、大量数据的分析、数值优化等），可考虑将计算任务分散到跨域的数据中心并行处理；而对于紧耦合式计算（例如迭代计算等）应集中调度到单一数据中心进行计算。例如，随着数据和参数规模的快速增长，将机器学习、深度学习模型进行分布式训练成为趋势。然而，由于分布式迭代训练过程往往需要传输大规模参数数据，需要综合考虑模型训练性能（延迟、模型质量等）和系统代价，探索紧耦合分布式参数服务器内存计算[2]模式。最后，可考虑对融合场景下计算任务并发调度的支持。例如，短临天气预报应用同时涉及科学计算、智能计算以及大数据处理。基于共享异构算力的一体化数据中心，可以将应用涉及的不同类型计算任务调度到合适的算力资源进行并行处理。

存储资源共享

存储资源共享和管理是一体化大数据中心资源共享的重要方面。存储管理应充分考虑用户的地理分布、计算资源本地性、应用数据的冷热程度和数据的访问模式等因素。跨域共享巨量多类存储资源是一体化大数据中心的潜在优势。应有机利用跨域海量异构存储资源，构建高速度、巨容量、低成本的存储系统，支撑上层应用的存储需求。应考虑如何针对用户、应用和所涉算力资源的位置特征、数据访问的模式特征等，构建充分考虑存储本地性的跨域分布式存储系统；考虑如何根据数据的冷热度，构建基于异构存储器件的多层次海量存储系统。例如，实际应用系统中的数据访问热度一般随时间增长而降低，可以针对近期频繁访问的热数据，构建基于分布式内存的大规模缓存系统，并将数据存储在靠近用户的区域节点上，以提升热数据的访问效率；相反，针对访问频率不高的冷数据，可充分发挥磁存储、光存储、玻璃存储介质容量大的特征，构建大规模冷数据存储系统。又如，在实际大数据应用中，对于多版本数据的快速存储和访问，可采用数据基础版本结合增量存储的方式进行组织，同时设计多层次异构混合存储系统，将读多写少且容量大的基础版本数据存储于持久内存中，而将频繁写的增量数据存储在普通内存系统中。

高性能可靠数据传输

高效网络传输 构建全国一体化中心涉及海量数据传输，包含单数据中心内部传输和跨域数据中心间传输。在数据中心内部，远程直接数据存取（Remote Direct Memory Access，RDMA）网络技术近年来得到快速发展。RDMA允许用户态的应用程序直接读取或写入远程内存，而无内核干预和内存拷贝发生，有效解决了网络传输中服务器端数据处理的延迟。长期以来，受传统网络条件限制，须遵循数据本地性调度计算任务，但这样会导致计算资源不可调度的问题，造成资源的严重浪费。RDMA网络将打破这一传统限制，突破系统可调度性瓶颈。随着软件定义网络（Software Defined Network，SDN）的兴起，主流云计算厂商，如亚马逊、微软和谷歌等，均在推进面向跨数据中心的软件定义广域网（SD-WAN）的建设。此外，国际互联网工程任务组（IETF）也制定了广域网范围基于TCP/IP的RDMA扩展协议，即iWARP[3]，将为跨数据中心数据高效传输提供机会。跨域数据中心间传输应支撑跨域大数据计算和存储，因此既要遵循标准化的网络传输协议以提高通用性，又要考虑大数据处理环境。例如，现有的大数据存储系统中一般充分考虑了并行化和容错需求，采用了分片存储技术[4]，在进行数据跨域处理时，可考虑支持远程大数据处理的数据部署“克隆”机制，以实现大数据远程处理环境的快速部署。为保障数据传输的安全可控，可考虑第三方传输技术。面向网络传输失效问题，须提供可靠的断点续传。针对超规模数据，须提供离线化安全无缝接入方法。

云边端融合近数处理 边缘计算是由海量地理分布式边缘服务器构成的网络边缘侧开放平台，随着5G、超5G、6G等通信技术的发展，边缘计算能够就近在用户侧提供计算、存储和网络资源等服务，缓解端侧算力不足而云侧时延高的矛盾，形成了云边一体化的计算架构。然而，区别于传统云计算中心相对稳定集中的资源供给方式，云边一体化平台具有硬件异构和资源受限等特征。

同时，云边平台中需要处理来自用户的多样化服务，亟须发展多资源协同调度的新理论与新方法，包括云边协同的任务卸载机制和边缘平台资源的高效匹配机制等。

随流检测智能管理网络 一体化数据中心通过骨干广域网连接多个云数据中心，实现用户、云中心间的高速互联。数据中心资源的无缝接入涉及大量网络流的传输，如何提高网络的智能运维能力、便捷化网络管理对用户接入云、提高服务质量至关重要。随流检测技术（in-situ Flow Information Telemetry, iFIT）通过对真实网络业务流报文进行性能测试，在减少对网络传输干预的前提下，提供真实逐跳网络时延、抖动、丢包等性能指标，为定位网络故障、智能管理网络提供必要信息。随着 SDN 的不断发展，网络测量应更加适应软硬件结合以及可编程环境，基于 sketch 的高速网络流量技术被赋予厚望，融入到 iFIT 中从而实现面向通用或特定任务的网络测量目的[1]。

数据和计算的安全保障

数据和计算安全是一体化数据中心的保障。在数据安全方面，通常在将数据存储到外部云提供商之前对其进行加密。然而，加密限制了提供商侧的数据查询等操作。这个问题一般通过两种方法解决：其一是定义索引，即在提供商侧进行（部分）查询评估，而无须解密数据；其二是使用加密技术，即在加密数据上直接执行操作或条件评估。定义索引需要平衡精度和隐私。支持加密数据执行操作的加密技术包括：允许评估范围条件的保序加密，以及允许在加密数据上评估任意复杂函数的全同态（或半同态）加密。以这些加密技术为基本构件，可以构建安全高效的加密数据库系统，以支持对加密数据的查询。

在计算安全方面，可采用基于硬件的 TEE 技术来保障提供商侧敏感计算的安全。TEE 技术通过软硬件方法构建隔离的运行环境，提供数据加密、隔离执行等功能。现有的支持 TEE 的硬件包括英特尔的软件保护扩展（Software Guard Extensions，SGX）和 AMD 的安全加密虚拟化（Secure Encrypted Virtualization，SEV）等。SGX 使用指令集扩展和访问控制机制来隔离程序运行环境，保护飞地的代码和数据不被恶意操作系统和虚拟机管理器访问。SEV 通过对虚拟机进行加密达到隔离虚拟机、防止特权软件攻击的目的，对虚拟机上的程序透明。但是，目前的 TEE 与 CPU 硬件紧密结合，缺乏对不同硬件平台的兼容性；同时，虽在功能性、安全性和可部署性方面不尽相同，但仅围绕 CPU 进行保护，需要研究有效支持 GPU 等加速器的异构计算系统。

支撑应用

一体化大数据中心的建设，将助力我国数字经济基础设施全面升级，为数字经济产业提供有力支撑，全面推进 5G、工业互联网、人工智能等产业应用快速发展。建设一体化数据中心在加快技术创新和供给的同时，也要加快拓展典型应用场景，加快下游用户的融入。在 5G 产业领域，中国作为技术的全球领跑者，专利申请数量占比全球领先，2022 年底我国 5G 基站总数将突破 200 万，5G 终端用户总数将占全球的 80% 以上。应结合我国 5G 网络建设的快速推进和领先优势，充分提升一体化数据中心的用户接入和云边端数据传输的能力，在应用层结合 5G 终端产品的开发，依托一体化数据中心加快开发和提供超高清、沉浸式等服务应用。在工业互联网领域，依托一体化数据中心，加快对传统产业链的融合。我国拥有全部产业门类且已连续 12 年保持世界第一制造大国地位，应基于一体化数据中心大力推进工业互联网建设，促进我国产业链跨越式升级。例如，利用一体化数据中心的丰富算力资源，构建工业数字孪生底座，实现制造与信息的深度融合，解决工业制造在设计、制造、调试、运行、维护全生命周期中的问题。在一体化数据中心的支撑下，工业互联网可以发挥共享基础算力的优

[1] 参见 http://www.watersprings.org/pub/id/draft-song-opsawg-ifit-framework-16.html。

势，也可以借助超算算力提高先进制造领域的效率，例如在大飞机机翼、汽车的轴承、发动机气缸等的疲劳试验中采用超算算力，可以有效节省开发周期，节约研发成本。

在人工智能应用领域，可考虑依托一体化数据中心对人工智能模型算法进行服务化重用。随着当前人工智能大模型的发展与普及，每重新训练一个大模型都需要大量计算资源。例如，训练一个 GPT-3 模型需要利用 10000 块 NVIDIA V100 GPU 同时训练 15 天，大约花费 120 万美元，产生 85000 千克二氧化碳。考虑对已训练好的模型进行复用具有重要意义，目前人工智能开源平台 Hugging Face[2] 已托管了 5 万个人工智能模型，并发展成市值 20 亿美元的公司。同时，人工智能模型算法的服务化重用有利于提升用户开发效率和使用体验，对一体化数据中心的推广具有较好的作用。此外，一体化数据中心也可为面向模型服务的集成智能提供更多低成本的算力资源。例如，可采用多模型训练集成的方式，聚合多个小模型，获得低耗时、高质量、低成本的复合模型 [5]。最后，服务化部署也可使底层细粒度资源差异对用户透明，这是实现平台跨域互操作的成功途径 [6]。

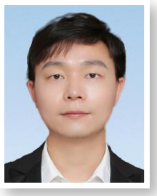
总结

建设全国一体化大数据中心将为加快我国数字经济的发展提供重要支撑。本文围绕建立全国一体化大数据中心的支撑系统，从跨域资源共享（包括异构计算资源共享和存储资源共享）、高性能可靠数据传输（包括高效网络传输、云边端融合近数处理、随流检测智能管理网络）、数据和计算的安全保障、支撑应用等四个方面进行了初步思考和探讨。构建全国一体化大数据中心，在关键技术上存在大量开放性课题值得探索，希望本文能抛砖引玉，引发更多对构建全国一体化大数据中心关键技术的讨论。 ■

[2] 参见 https://huggingface.com。

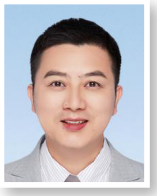
陈汉华

CCF 专业会员，CCF 传播工委委员。华中科技大学计算机学院教授。主要研究方向为分布式计算系统、大数据处理系统。
chen@hust.edu.cn



王　雄

CCF 专业会员，CCF 抗恶劣环境计算机专委会委员。华中科技大学计算机学院副教授。主要研究方向为分布式学习系统、联邦学习、网络流控 / 拥塞控制云计算与边缘计算。xiongwang@hust.edu.cn



华强胜

CCF 高级会员，CCF 武汉秘书长。华中科技大学计算机学院研究员。主要研究方向为并行分布式计算理论与算法。
qshua@hust.edu.cn

其他作者：顾　琳　羌卫中　金　海

参考文献

[1] Buyya R, Srirama S N, Casale G, et al. A Manifesto for Future Generation Cloud Computing: Research Directions for the Next Decade[J]. *ACM Computing Survey*, 2019(5):105:1-105:38.

[2] Zaharia M, Xin R S, Wendell P, Das T, et al. Apache Spark: A unified engine for big data processing[J]. *Communications of the ACM*, 2016(11): 56-65.

[3] IETF RFC 4296, The Architecture of Direct Data Placement (DDP) and Remote Direct Memory Access (RDMA) on Internet Protocols[S/OL]. (2005-12). https://www.hjp.at/doc/rfc/rfc4296.html.

[4] Shvachko K, Kuang H, Radia S, et al. The Hadoop Distributed File System[C]// *Proceedings of the IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST 2010)*, 2010:1-10.

[5] Gunasekaran J R, Mishra C S, Thinakaran P, et al. Cocktail: A Multidimensional Optimization for Model Serving in Cloud [C]// *Proceedings of the 19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2022)*, 2022:4-6.

[6] Foster I T, Kesselman C. The History of the Grid[OL]. arXiv:2204.04312, 2022.