面向国家高性能计算环境的虚拟数据空间系统

秦广军[1], 肖利民[2,3], 张广艳[4], 牛北方[5,6], 陈志广[7]
1. 北京联合大学智慧城市学院, 北京 100101; 2. 北京航空航天大学计算机学院, 北京 100191;
3. 软件开发环境国家重点实验室, 北京 100191; 4. 清华大学计算机科学与技术系, 北京 100084;
5. 中国科学院计算机网络信息中心, 北京 100190; 6. 中国科学院大学, 北京 100190;
7. 中山大学计算机学院, 广东 广州 510006

摘要
高性能计算环境是支撑国家科技创新、经济发展、国防建设的核心信息基础设施，世界高性能计算强国纷纷建设基于多超算中心资源的广域高性能计算环境。然而，高性能计算环境中资源种类繁多且地域分布广，无法有效发挥资源的聚合效应，难以满足大型应用对广域分布数据的统一管理和高效访问需求。为此，提出了一套可用于构建广域全局虚拟数据空间的完整技术体系，包括虚拟数据空间模型、跨域虚拟数据空间构建、广域环境中数据高效迁移、广域环境中存算协同调度、跨域高并发数据聚合处理等技术，并研发了一个可运行于国家高性能计算环境的虚拟数据空间系统，可有效支撑广域分散异构存储资源的统一高效访问，实现广域环境中分布数据的跨域共享和协同处理。目前，该软件系统已在国家高性能计算环境实验性部署，并验证了分子对接、全基因组关联分析、天气预报模式3类典型大型应用。验证结果表明，所研虚拟数据空间构建方法和系统可有效聚合广域分散的存储资源，满足大型应用的数据空间需求。

关键词
高性能计算环境；大型计算问题；虚拟数据空间；广域分布式存储；统一命名空间

中图分类号：TP316 文献标识码：A doi: 10.11959/j.issn.2096-0271.2021016

Virtual data space system for national high-performance computing environment

QIN Guangjun[1], XIAO Limin[2,3], ZHANG Guangyan[4], NIU Beifang[5,6], CHEN Zhiguang[7]
1. Smart City College, Beijing Union University, Beijing 100101, China
2. School of Computer Science and Engineering, Beihang University, Beijing 100191, China
3. State Key Laboratory of Software Development Environment, Beijing 100191, China
4. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
5. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China
6. University of Chinese Academy of Sciences, Beijing 100190, China
7. School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China

Abstract
High-performance computing (HPC) environment is the core information infrastructure supporting national scientific and technological innovation, economic development and national defense construction. High-performance computing powers

around the world have been building wide-area HPC environments based on multi-supercomputing center resources. However, in the high-performance computing environment, there are many kinds of resources and wide geographical distribution, which cannot effectively exert the aggregation effect of resources, and it is difficult to meet the requirements of large-scale applications for unified management and efficient access to wide-area distributed data. To this end, a complete set of technologies were proposed, which could be used to build wide-area global virtual data space, including virtual data space model, cross-domain virtual data space constructing, efficiently migrating data in a wide-area environment, co-scheduling of storage resources and computing job and cross-domain high concurrency data aggregation processing, etc. Based on the above, a virtual data space system has been developed for the national high-performance computing environment (NHPCE), which can effectively support the unified and efficient access to the wide area distributed heterogeneous storage resources, and the distributed data in the wide-area environment can be shared and cooperative processed in a cross-domain manner. At present, the system was experimental deployed in NHPCE and three typical large-scale applications, such as molecular docking, genome-wide association study and weather forecasting model, have been verified. The verification results show that the developed technology and software system can effectively aggregate the wide area distributed storage resources and meet the data space requirements of large-scale applications.

Key words
high-performance computing environment, large-scale computing problem, virtual data space, wide-area distributed storage, unified namespace

1 引言

高性能计算环境是支撑国家科技创新、经济发展、国防建设的核心信息基础设施，世界高性能计算强国纷纷建设基于多超级计算中心（以下简称超算中心）资源的广域高性能计算环境[1]。美国建立了跨域的极限科学与工程发现环境（extreme science and engineering discovery environment，XSEDE），旨在建设单一的虚拟系统，世界各地的科学家可以通过系统共享计算资源、数据和专业知识；欧洲建立了跨域的欧洲网格基础设施（European grid infrastructure，EGI），目的是通过整合数字功能、各界资源和专业知识为科学研究和基础设施建设提供开放的解决方案；我国建立了中国国家网格（China national grid，CNGrid），通过资源共享、协同工作和服务机制，有效地支持科学研究、资源环境、先进制造和信息服务等应用。高性能计算水平体现了一个国家的科技综合实力，整合广域分散的高性能计算资源，建立广域高性能计算环境，对于国家高性能计算技术的领先发展、国家安全与高性能计算地位的提高至关重要。

与XSEDE和EGI相比，CNGrid不仅要能够支持科学研究，更强调对多领域应用的支持。这些应用除了需要高性能计算能力，还需要支持对异地、异构数据进行存储、访问、交换和处理的能力。然而，在广域高性能计算环境中，各超算中心往往地理位置分散，资源自治管理，数据跨域分散存储，这使得资源和数据难以统一管理、调度和互访，应用系统间相互孤立，难以满足大型计算应用对全局资源空间的需求。因此，如何在广域高性能计算环境中实现跨域资源统一管理与使用，有效支撑大型计算应用，一直是各高性能计算领域的重要

研究课题，这迫切需要新技术、新系统来支持资源共享，提高资源利用率，发挥分散资源聚合效应。

CNGrid目前已经支持全局计算资源管理和作业调度，但存储和数据资源仍然不能得到有效的全局统一管理、调度和访问。本文针对国家高性能计算环境广域分散存储资源的聚合需求及大型计算应用对跨域全局虚拟数据空间的实际需要，对标高性能计算环境广域存储系统EGI OneData[2-3]和XSEDE GFFS[4]，从跨域虚拟数据空间构建、广域数据共享、全局存算协同调度、跨域并发数据聚合处理、CNGrid环境对接等几个主要方面出发，建立了一套可用于构建广域全局虚拟数据空间的完整技术体系，并研发了一个可运行于国家高性能计算环境的虚拟数据空间系统，旨在为在国家高性能计算环境中建立虚拟数据空间提供技术手段、应用经验、人才储备，支撑建设资源共享、统一管理、高效协同的国家高性能计算环境，促进我国高性能计算环境的应用和可持续发展。

2 国内外研究现状

国家级广域高性能计算环境是支撑国家科技创新、经济发展、国防建设的核心信息基础设施，是大国竞争的战略高地，世界高性能计算强国纷纷建设基于多超算中心资源的广域高性能计算环境。

美国、欧洲、日本对虚拟数据空间系统及关键技术开展了研究。美国国家科学基金会的TeraGrid计划[5]及其后续的XSEDE计划[6]，以及欧洲的网络基础项目EGI（前身为EGEE）[7]，都旨在将广域分散自治的大规模计算系统、科学仪器等互连并广域共享，但TeraGrid需采用专用高速网络，EGI欠缺全局统一管理能力。其中，EGI的基础存储系统是OneData，引入了“空间”和“供给者”的概念，较好地屏蔽了EGI中数据广域分布的复杂性，但是采用紧密的元数据管理方式，元数据维护压力巨大，系统可扩展性较差。XSEDE的基础存储系统是全局联合文件系统（global federated file system，GFFS），采用松散的顶层元数据组织实现了异构存储资源的聚合，但是顶层元数据集中管理，存储集群的元数据分散自治管理，使得顶层元数据极易成为性能瓶颈。麻省理工学院的协作式文件系统（cooperative file system，CFS）、加利福尼亚大学伯克利分校的OceanStore[8]、纽约大学的Kademlia[9]等具有良好的平衡性和扩展性，但均为聚合集中式存储资源的系统。谷歌公司的Spanner[10]实现了在特定硬件支撑下的跨域数据库存储模式，耶鲁大学和谷歌公司联合实现了跨数据中心的CalvinFS系统[11]，加利福尼亚大学河滨分校提出了可跨多云平台的SPANStore系统[12]，德国卡尔斯鲁厄理工学院设计了MetaStorage系统[13]，上述系统可管理分散的存储资源，但主要面向互联网应用（如数据库存储），不适用于高性能计算应用环境。微软公司的WAS（Windows Azure storage）系统[14]通过位置服务器和全局命名空间整合跨域存储集群，但不支持跨域数据共享。美国印第安纳大学实现了跨域的Lustre-WAN文件系统[15]，但需专用网络支持。日本筑波大学提出了跨域网格文件系统Gfarm[16]，但其集中式元数据架构难以适应高性能计算环境的大规模并发数据访问请求。

我国对虚拟数据空间系统及相关技术也开展了相关研究，建设了基于多个超算中心的国家高性能计算环境，实现了分散计算资源的统一管理和全局调度，但尚未实现分散存储资源的全局数据空间以及存

储与计算全局协同调度。电子科技大学、中国科学院计算技术研究所、浙江大学等采用哈希算法，设计了针对集中式存储资源的聚合系统PeerStore[17]和π-Store等。北京邮电大学、华为技术有限公司、阿里巴巴集团[18]面向互联网应用实现了基于多云存储平台协同的云存储模式。清华大学[19]、北京航空航天大学[20]、中国科学院计算机网络信息中心研究了单一大规模存储聚合系统及跨域存储聚合技术，针对分布性、异构性、动态性的广域网络环境，实现了支持跨域数据驱动型应用的虚拟数据空间及服务协同平台、跨多数据中心的全局虚拟文件系统等。

综上，目前国内外都在研究跨域存储资源聚合、广域数据共享等问题，但尚未出现可有效支持广域高性能计算环境的跨域虚拟数据空间。因此，研究在广域高性能计算环境中建立跨域虚拟数据空间的方法和关键技术具有重要的理论意义和应用价值。

3 国家高性能计算环境

3.1 环境现状

我国国家高性能计算环境（原中国国家网格环境）的环境资源种类繁多、异构性强、地域分布广，主要由上海超级计算中心和中国科学院计算机网络信息中心两个南北主节点，国家超级计算无锡中心、国家超级计算天津中心等7个国家级超算中心，以及清华大学、西安交通大学等11个普通节点组成，总计算能力超200 PFlops，总存储容量超160 PB，2020年新增国家超级计算郑州中心和国家超级计算昆山中心。

计算资源管理的核心系统软件——超级计算环境（super computing environment，SCE）[21]是中国科学院开发的环境中间件，用户可以通过此中间件使用整个环境中的所有计算资源。SCE主要包括前端服务器（front server，FS）和中央服务器（center server，CS），CS负责汇总FS采集的各类信息，以及作业全局调度与管理服务、数据传输与管理服务、用户与权限服务、资源信息管理服务、安全策略以及计算环境管理；FS负责资源接入与监控、作业局部调度、局部信息管理、一些计算资源的执行控制，收集来自各超算中心的资源信息，并汇报给CS，以及执行来自CS的各种执行请求。

存储资源由各超算中心自治管理，使用方式主要分为两类：第一类，在超算中心中将区域划分为计算区和存储区，采用不同的文件系统进行管理并存储在不同的集群上，进行计算作业时，需要将用户的作业及用到的数据提交至计算集群中进行计算；第二类是不划分存储区和计算区，存储和计算由同一个文件系统统一管理，作业直接在用户目录下运行。对于当前两种使用模式，用户数据都汇聚在一个超算中心中，而且是分散自治的，国家高性能计算环境中各超算中心之间相互隔离，无法做到用户数据跨域及统一管理。

可见，在当前的国家高性能计算环境中，计算资源可统一管理、全局调度，但存储资源仍广域分散、隔离自治，虽然可全局调度计算资源，但无法有效地实现数据的跨广域统一访问和共享，应用规模的扩展受限于单中心的资源规模，无法构建更大型的、跨广域的应用，更无法实现存储与计算的协同调度，从而导致全系统资源利用率不能有效提高。

3.2 大型计算问题对数据空间的需求

大型计算问题，诸如生物信息、精准医疗、高能物理、气象预报等类型的应用，

由于数据量和计算量都较大[22]，且数据往往跨广域分布，需要在高性能计算环境中形成广域的数据共享、统一的数据空间，从而提高应用的规模，提高全系统资源利用率。例如，生物信息和精准医疗类应用涉及的数据量巨大，存储需求往往达PB级，且需要在跨广域海量样本中进行汇聚处理和挖掘，而单中心局部存储空间不足以满足应用需求，且受到广域网带宽和路由的限制，数据跨广域迁移效率较低，可统一管理的存储空间和高效的广域数据共享将有利于此类应用在数据处理规模上的扩展；高能物理类应用往往需要E级计算，目前单中心的计算能力尚不能有效满足其需求，人为设置的数据和任务布局并不能很好地依据各中心具体的资源提供能力进行优化，实现数据与计算任务的协同布局和调度将有利于此类应用的高效运行；气象预报类应用涉及广泛的数据源，其类型多、分布广，且时效性要求高，同样受到广域网带宽和路由的限制，数据跨域访问性能较低，实现跨广域的多源数据聚合处理也将有利于提高此类应用的性能和时效性。

具体来讲，数据空间应满足如下4个方面大型应用的要求。

（1）支持跨域存储资源统一管理和访问

由于国家高性能计算环境中存储资源广域分散且隔离自治，系统无法对跨域存储资源进行有效的管理和统一访问，导致各数据中心数据的重复存储以及多超算中心无法协同处理数据。大型应用迫切需要将分散的存储资源聚合为全局数据空间，并提供跨域统一管理和访问能力。

（2）支持广域数据共享

在高性能计算环境中广域数据无法共享，这导致资源闲置、重复建设，数据空间需要提供能够有效整合分散自治、广域隔离的存储资源，汇聚各超算中心开放的数据，为用户提供跨多个节点提取数据的能力，为应用提供一站式的数据共享服务。

（3）支持存储与计算协同调度

由于高性能计算环境中的存储与计算无法高效协同，广域范围内的计算任务和数据难以实现合理分布，需要设计并开发虚拟数据空间与现有国家高性能计算环境软件的接口，以支持虚拟数据空间与国家高性能计算环境的对接，通过技术集成形成虚拟数据空间系统，提供能够透明实现多节点聚合的机制，且能根据计算特征和数据布局来控制任务与数据的节点选择，从而实现存算协同。

（4）支持跨域多源数据聚合处理

为了提升典型数据访问模式的跨域访问能力，需要能对数据空间与应用之间的I/O中间层进行优化、对跨域多源高并发数据进行高效聚合处理的方法来有效支持大型计算应用。因此，大型计算问题亟须聚合广域分散的存储资源，形成跨域的高性能计算数据空间，以满足大型计算问题在规模、性能和资源利用率上的要求。

此外，从使用者角度考虑，还应该具备如下跨域分布式存储系统的基本功能。

- 可跨广域环境进行基本的存储操作，且符合文件系统的标准可移植操作系统接口（protable operating system interface of UNIX, POSIX协议），如重命名、修改、增加、删除等。
- 可访问广域环境中的数据集及其子集，且符合文件系统的标准POSIX协议。
- 可在广域环境中统一浏览文件目录，且有权限限制。
- 可将数据共享给指定的其他或所有用户。
- 可将单个或多个数据源数据按需存储到一个或多个超算中心，且能统一浏览和访问。
- 可依据数据访问特性进行优化的全

局资源分配与数据放置。

4 高性能计算虚拟数据空间

针对上述大型计算问题对数据空间的要求，笔者在跨域虚拟数据空间的模型、体系结构、资源聚合、管理和访问等方面开展了研究，突破了广域分散自治存储资源聚合、带宽约束下数据高效可靠迁移、计算与数据跨域协同调度、高并发数据流聚合处理等关键技术，形成了完整的虚拟数据空间系统，并已经在国家高性能计算环境中对该系统进行了初步部署和应用验证。

4.1 虚拟数据空间体系结构及关键技术

4.1.1 虚拟数据空间模型

针对广域分散、自治异构的底层存储资源，以及种类繁多、需求各异的上层大型应用，通过分析数据分布需求和应用的数据访问特征，抽象底层存储资源的分布形式，将虚拟数据空间提炼为主体、服务、空间、资源四要素，并构建了多层级的数据空间模型，如图1所示。

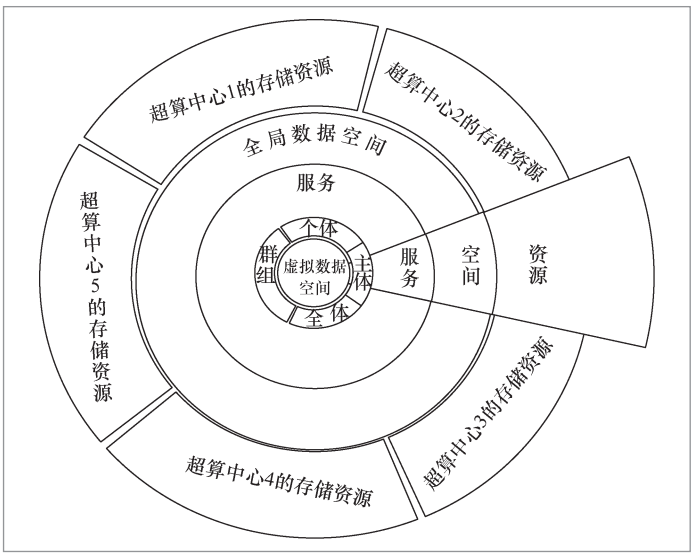


图1 虚拟数据空间模型

各要素的定义及形式化描述如下。

- 主体：数据空间服务的对象（如个体用户、群组用户、全体用户及其应用），主体可通过服务的形式使用虚拟数据空间的资源。
- 服务：面向主体提供的各种功能服务，如用户管理、区域管理、空间管理、权限管理、访问控制、数据共享、数据迁移等。
- 空间：对分散自治的存储资源依次进行物理聚合、局部聚合、全局聚合形成的全局虚拟数据空间。
- 资源：广域分散、隔离自治的存储资源。

4.1.2 虚拟数据空间表示方法

基于上述虚拟数据空间模型，对虚拟数据空间进行层次化表示，形成虚拟数据空间的层次化模型，主要包括资源层、空间层、服务层、主体层，如图2所示。

- 资源层：包含各超算中心的存储资源，存储资源分布在不同的地理位置上，且通常具有异构性。
- 空间层：通过对底层广域分散的存储资源依次采用物理存储资源聚合、局部存储资源聚合、全局存储资源聚合，最终形成全局虚拟数据空间。
- 服务层：提供使用虚拟数据空间存储资源所需的基本功能，主要包括用户管理、区域管理、空间管理、权限管理、访问控制、数据共享、数据迁移等，并通过统一接口以服务形式对外提供。
- 主体层：主要包含用户及其应用（如数值模拟、大数据、人工智能等典型应用），可通过接口使用虚拟数据空间提供的各种服务。

4.1.3 虚拟数据空间软件体系结构

基于本文提出的模型和表示方式，将

虚拟数据空间体系结构相应地设计为资源层、空间层、服务层、主体层4个层次，如**图3**所示。

● 资源层：处于最底层，主要提供用于构建虚拟数据空间的物理存储资源。该层包含广域分布的存储资源，存储资源分布于不同地理位置的超算中心之中。

● 空间层：位于资源层之上，通过聚合底层广域分散的存储资源，形成全局数据空间。该层依次采用物理存储资源聚合、局部存储资源聚合、全局存储资源聚合等资源聚合方法，实现广域存储资源的逐层聚合；同时，采用全局名字空间节点高可用方法实现全局元数据关键组件的高可用。

● 服务层：提供虚拟数据空间基本服务，如数据区域划分和管理提供按需区域划分服务；区域空间分配和管理提供区域映射和空间分配服务；区域隔离和权限管控提供区域隔离和数据安全保障；数据访问优化通过元数据访问优化和远程数据缓存提升元数据和数据的访问性能；数据访问带宽聚合服务用来优化频繁访问数据的广域布局，以提高带宽利用率；数据迁移共享通过优化应用I/O与迁移速率、多源与多数据迁移性能提高数据迁移与共享效能；安全可靠传输机制提供构造可靠迁移协议和高效安全迁移服务；存算协同调度提供数据传输、放置及任务布局协同的全局作业调度服务；访问接口服务为不同应用对虚拟数据空间的统一访问提供命令行和文件视图两种接口使用方式，并提供数据聚合处理框架和并行I/O库，以优化大型应用常用的高级I/O接口，并支持复杂数据处理模式，提升数据访问性能。

● 主体层：主要包含各类用户及其应用，如天气预报模式、生物信息学、目标协同识别等典型应用，应用可通过调用服务层提供的服务功能，在全局虚拟数据空间中使用资源层中的广域分布存储资源。

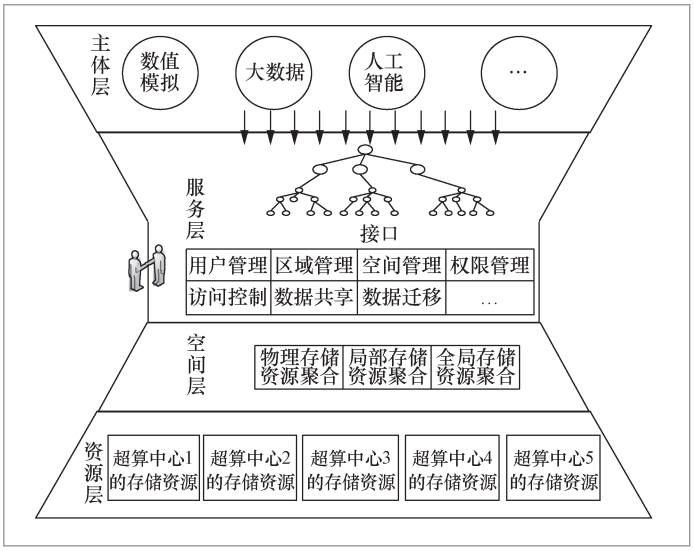


图 2 虚拟数据空间的层次化表示

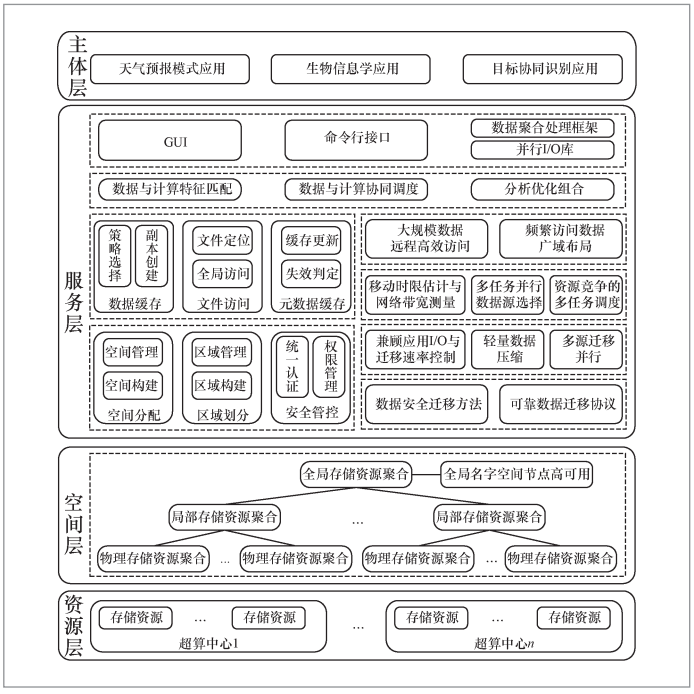


图 3 虚拟数据空间体系结构

4.1.4 虚拟数据空间系统关键技术

在本文建立的虚拟数据空间理论和模型的指导下，重点从跨域虚拟数据空间模型及构建方法、虚拟数据空间中数据的共享与迁移方法、国家高性能计算

环境中的虚拟数据空间运行支撑技术、面向典型应用的虚拟数据空间验证与优化技术4个层面开展研发工作，研发了一套面向高性能计算环境的广域数据存储与共享的技术体系和功能体系，具体涉及的关键技术如**图4**所示。

（1）跨域虚拟数据空间构建方法

针对广域分散存储资源的统一管理和高效访问需求，笔者依据虚拟数据空间理论模型，将广域分散自治的存储资源抽象为层次化模型，从本地、局域、广域3个层级进行聚合，构建与本地数据空间一致且能可靠地统一访问与管理的跨域虚拟数据空间，并定制化个人、群组、全局多级数据分区安全可靠共享，优化跨域元数据与数据服务能力，以解决跨域分散存储资源的统一管理和高效访问问题，有效发挥资源聚合效应。相比国外同类典型系统的相关技术，本文的跨域虚拟数据空间构建方法比OneData增加了管理数据的高可用能力，比Gfarm增加了管理数据高可用和数据区域划分能力，比CalvinFS增加了数据区域划分、跨域数据共享和账号安全管理能力。在性能测试中，基于本文方法构建的跨域资源聚合层软件模块在聚合访问本地单设备存储资源、本地单超算中心局部存储资源和跨广域全局存储资源方面，分别可达到直接访问存储资源时性能的96%、86.73%和84.3%；客户端元数据时延比基于最近最少使用（least recently used，LRU）的替换策略、基于目录（directory-directed prefetching，DDP）的预取策略、基于概率图（variant probability graph，VPG）的预取策略、基于Apriori关联规则算法的预取策略和基于语义距离算法的预取策略分别减少27.8%、32.5%、19.37%、24.96%、22.17%的平均访问时间[23]；维护数据一致性的开销比Raft-log减少42 ms；每秒查询数（queries-per-second，QPS）比Raft-log提升36倍，也优于MaterSlave和Tintri等系统，并且通过数据副本的优化布局，本地副本命中率达到68%，远程副本访问率下降至32%[24]。

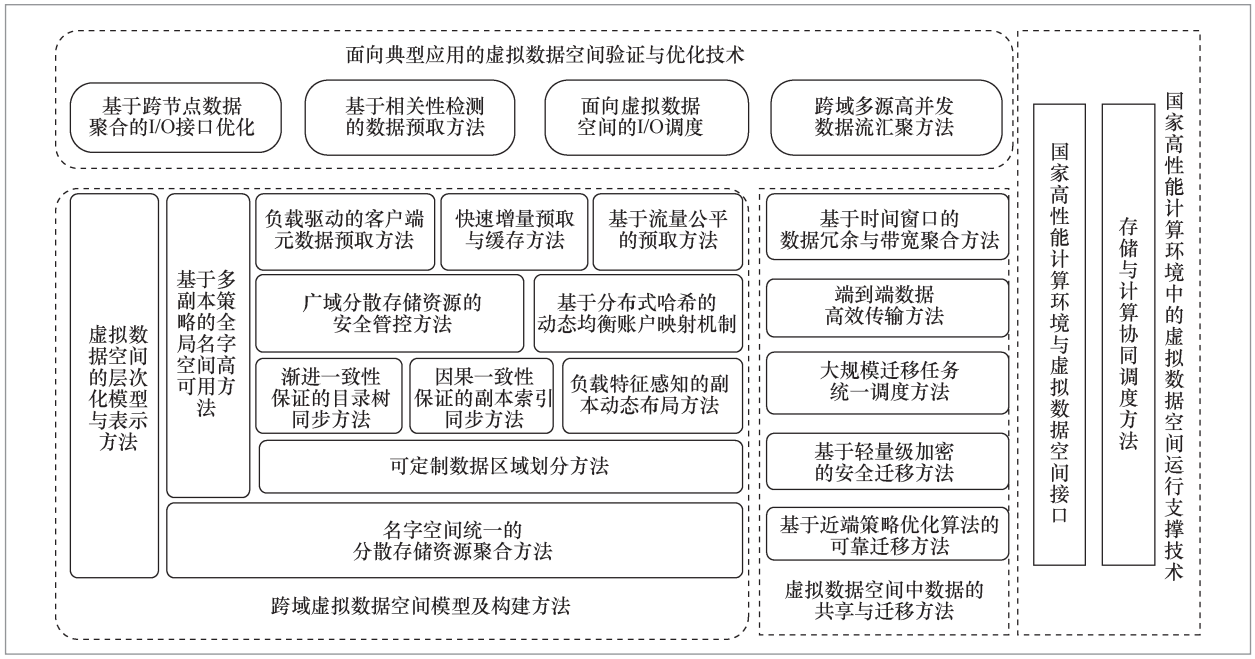


图 4 高性能计算虚拟数据空间关键技术

（2）广域环境中的数据高效迁移方法

针对大规模高性能计算数据跨域迁移中的带宽受限问题，笔者研发了可突破广域带宽受限的数据高效迁移方法，将网络拥塞控制过程抽象为可部分观察的马尔可夫决策过程，以动态凸包和迭代加权混洗方法来决策多迁移任务的调度和传输带宽分配，以多TCP流、流水线、并发传输等形式迁移数据，满足了广域带宽受限下的数据高效迁移需求，可有效跨域迁移数据。相比国外同类典型系统的相关技术，本文的广域环境中的数据高效迁移方法基于广域非专用网络，比Gfarm和CalvinFS增加了数据可靠安全迁移能力，比GPFS增加了数据区域划分和多副本能力；相比于盘古系统，本文方法的跨广域数据迁移性能提升2.96倍。实验表明，文件越大，本文方法的传输性能越好，小文件的传输性能也不低于网络传输性能的35%，且网络吞吐量可提高两倍以上。

（3）广域环境中的存算协同调度方法

针对广域环境中计算任务与存储资源的协同调度需求，笔者研发了广域环境中计算任务与存储资源的联动调度方法，将各中心资源聚合为虚拟队列，按计算任务和数据分布情况、集群队列排队情况进行归一化，并根据时间成本来决策计算作业与存储资源的协同调度，以有效发挥计算与存储资源的联动效应。实验结果表明，本文的广域环境中的存算协同调度方法可有效地提升资源使用率和计算作业的调度性能[25]。相比国外同类典型系统，本文方法创造性地提供了高性能计算环境中存储和计算资源的协同调度及布局能力。

（4）跨域高并发数据聚合处理技术

针对虚拟数据空间中的资源异构、数据流高并发且多源等特征，笔者研发了面向跨域高并发数据流模式的数据聚合处理技术，以代理方式跨域访问元数据，以高并发异步乱序数据流的细粒度任务调度形式汇聚跨域多源高并发数据流，优化面向跨域环境的I/O接口，满足了跨域作业的高效执行需求，可有效发挥虚拟数据空间对大型应用跨域运行的支撑能力。相比国外同类典型系统，上层应用可基于本文提供的多副本和广域环境中的存算协同调度能力，跨域高并发访问多源数据，实现数据聚合处理。相较于相关技术，上层应用的远程数据请求率可减少38%~71%，命中率比自适应替换缓存（adjustable replacement cache，ARC）和预取方法提升20.7%和28.8%，文件创建的吞吐率提升17%~93%，执行时间减少37%。

4.2 虚拟数据空间系统

4.2.1 虚拟数据空间系统架构

在关键技术研究的基础上，笔者研发了面向高性能计算的虚拟数据空间系统GVDS，技术上覆盖了虚拟数据空间模型和体系结构、分散资源聚合方法、端到端数据传输方法、存储与计算协同调度方法、跨节点数据聚合的I/O接口优化等20多项关键技术，功能上覆盖了全局数据空间、跨域数据存储、数据区域划分、多副本等10多项重要功能。系统的总体架构如**图5**所示。

笔者设计了高性能计算虚拟数据空间系统的操作界面，包括Web和命令行，如**图6**所示。

Web界面展示了运算时间、输入输出带宽、容量信息、服务器、节点数量、用户数量、总体容量等信息，以及部署的节点分布情况。命令行界面包括41条命令，如账户注册命令、区域注册命令、空间映射增加命令、管理员审批命令、用户区域查询命令等，命令的接口见**表1**。

相较国际同类领先系统，GVDS具备

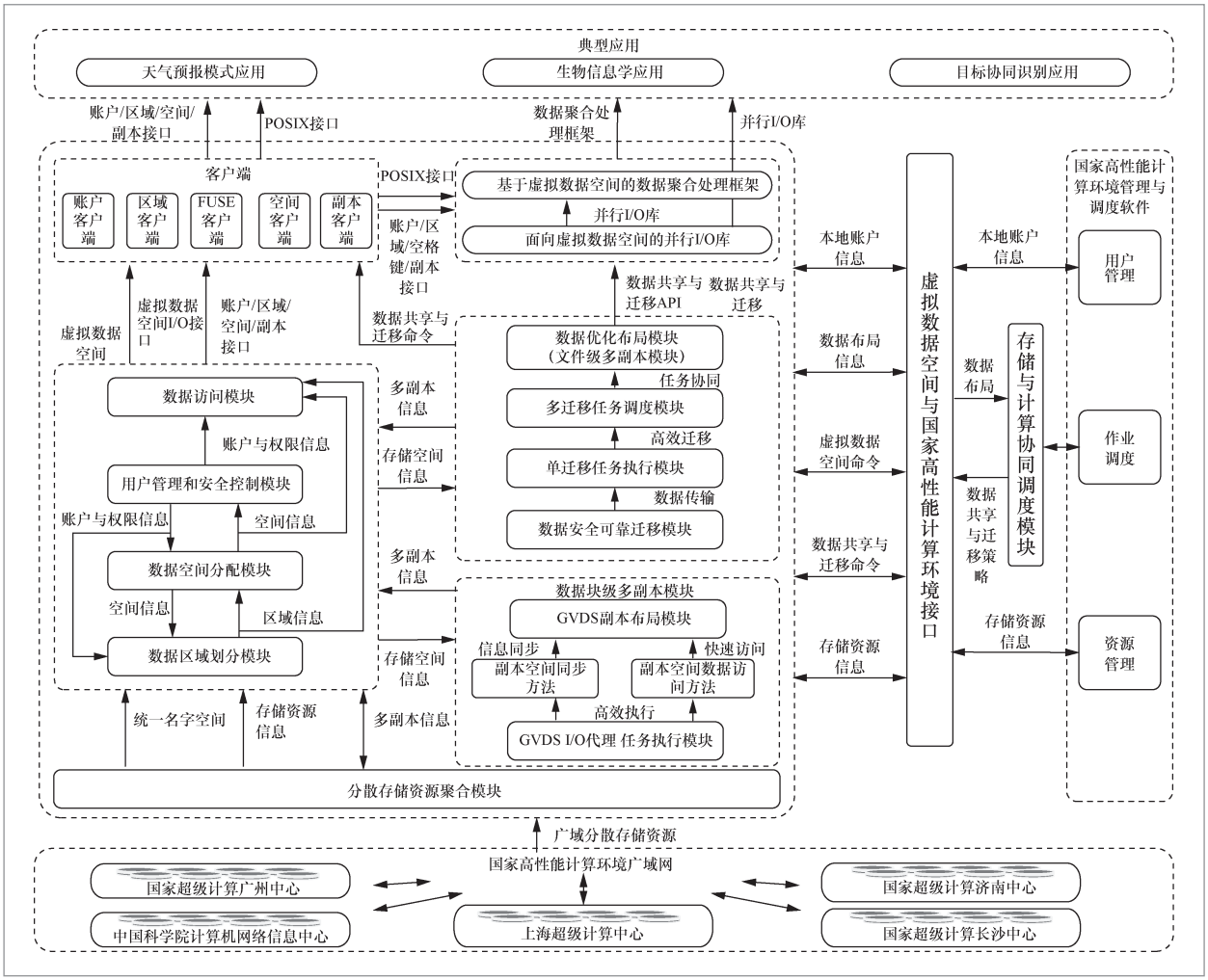


图 5 虚拟数据空间系统的总体架构

(a) 命令行界面 (b) Web界面

图 6 虚拟数据空间系统界面

表1 命令接口

命令	功能	命令	功能
usersignup	用户注册	Adspdeduct	管理员空间映射删除
userlogin	用户登录	Adzlist	管理员查询区域信息
userexit	用户登出	Adauthsearch	管理员权限信息查询
usercancel	注销用户	Adsprename	管理员空间重命名
usermodify	用户信息修改	Adzrename	管理员区域重命名
usersearch	用户信息查询	Adcam	管理员添加账户映射
zregister	区域注册	Adspusage	管理员查询空间已用容量
zrename	区域重命名	Adzshare	管理员区域共享
zcancel	区域注销	Adminsignup	管理员账户注册
zlist	区域信息查询	Aduam	管理员删除账户映射
zshare	区域共享	Adzsharecancel	管理员区域共享取消
spadd	空间映射增加	Adzadd	管理员区域增加
spdeduct	空间映射删除	Adseepool	管理员查看账户池
sprename	空间重命名	Adzcancel	管理员区域注销
spsizechange	空间容量修改	Authsearch	权限查询
spusage	空间已使用容量	Listapply	管理员查询请求信息
adauthmodify	管理员权限修改	Centerdelete	超算中心删除
suggestion	管理员审批请求信息	Rd	存储资源删除
centermodify	超算中心信息修改	Rg	存储资源注册
centersearch	超算中心信息查询	Rq	存储资源查询
ru	存储资源更新		

更完整的技术体系和功能。在技术体系上，该系统覆盖了数据空间模型、跨域空间构建、广域数据共享等20项关键技术，形成了完整的技术体系；在核心功能上，该系统与对标系统相比，有所超越，涵盖了全局名字空间、跨域数据共享、多数据副本等10项重要功能，形成了完整的功能体系，见**表2**。

所实现的虚拟数据空间符合文件系统的POSIX标准，可通过mount命令直接挂载，并支持多种异构文件系统，目前测试通过的文件系统包括Lustre、Ceph、GPFS、Gluster、MooseFS、ParaStore等符合POSIX标准的文件系统。目前，GVDS已在国家高性能计算环境的6个广域节点

表 2 GVDS 与国际同类系统的功能对比

功能	GVDS	GFFS	OneData	Gfarm	CalvinFS
全局名字空间	√	√	√	√	√
管理数据高可用	√	√	×	×	√
数据区域划分	√	×	√	×	×
跨域数据存储	√	√	√	√	√
跨域数据共享	√	√	√	√	×
统一访问接口	√	√	√	√	√
数据安全可靠迁移	√	√	√	×	×
账户与安全管理	√	√	√	√	×
缓存功能	√	√	√	√	√
多数据副本	√	×	√	√	√

上部署，可管理PB级的跨域存储资源，并在典型计算应用上进行了示范应用。初步测试表明，在关键性能上，与对标系统相比，该系统具有较大优势，跨域写数据和读数据性能分别是对标系统的1.3倍和1.6倍。

4.2.2 国家高性能计算环境部署

目前，笔者研发的系统已和国家高性能计算环境初步对接，虚拟数据空间与计算环境的访问接口采用REST风格的API，提供基于HTTP的国家高性能计算环境访问接口，包括集群节点、环境应用、环境队列等接口，为计算服务平台中的各类服务提供虚拟数据空间数据的查询、访问和传输。部署环境包括3个国家超级计算中心（国家超级计算广州中心、国家超级计算济南中心、国家超级计算长沙中心）、两个国家网格主节点（中国科学院计算机网络信息中心（中国国家网格北方主节点）、上海超级计算中心（中国国家网格南方主节点））。另外，也在北京航空航天大学完成部署，形成了跨广域6个节点测试验证环境。部署情况如图7所示。

所有节点各部署一个管理节点，2~3个I/O代理节点，以及一套Lustre文件系统。目前，所部署的验证环境已经汇聚1.57 PB存储空间，汇聚的各中心资源见表3。

目前，所研发的Web界面也被集成到中国国家网格门户网站“聚合资源运行支撑环境”AROSE平台中，可以通过AROSE

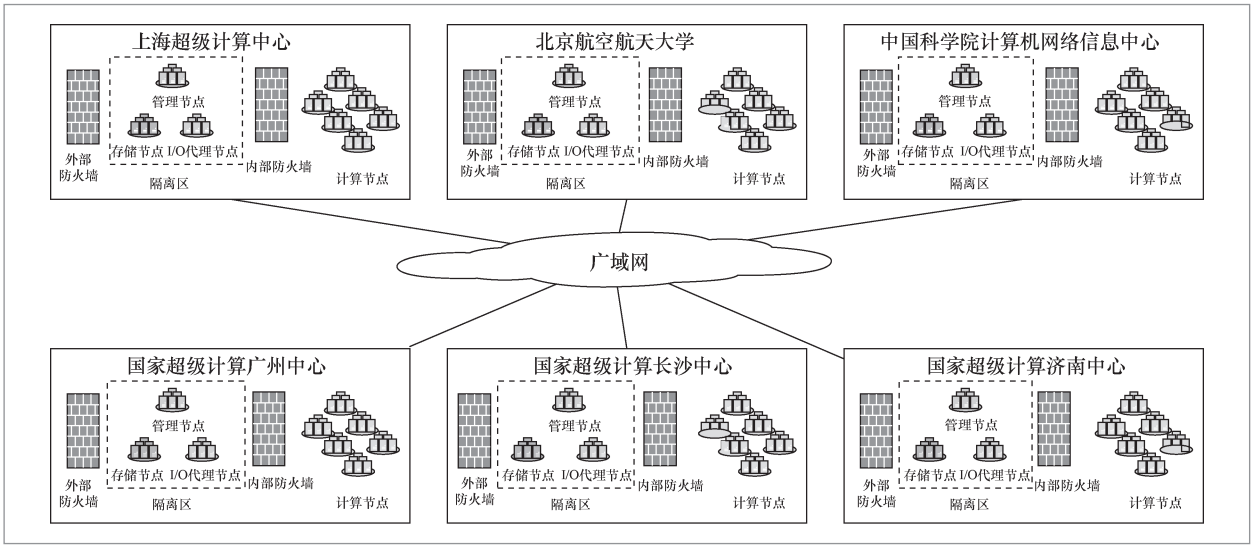


图7 系统部署情况

表3 目前部署环境已聚合的存储资源

节点	核数/个	存储量
中国科学院计算机网络信息中心	112	400 TB
上海超级计算中心	136	2 TB
国家超级计算广州中心	192	300 TB
国家超级计算济南中心	148	282.6 TB
国家超级计算长沙中心	64	96 TB
北京航空航天大学	432	489.4 TB
合计	1 084	1.57 PB

平台进入虚拟数据空间系统的Web界面。AROSE平台集成如图8所示。

4.3 典型应用验证

为了验证虚拟数据空间对应用的支撑效果，笔者在实验床上开展了典型场景和应用的测试验证工作。典型场景包括数据区域的定制化共享、远程大数据集的按需随机访问、广域分布数据的多中心协同处理、工作流作业的透明数据处理4类，典型应用包括生物信息学应用、跨域目标协同识别、天气预报模式等。验证方案如图9所示，主要验证全局统一视图、存储计算协同、广域数据共享等重要的特色功能。

截至目前，已经验证了生物信息学方面的分子对接应用、全基因组关联分析应用，以及天气预报模式应用，跨域目标协同识别应用还在部署中。具体应用情况如下。

（1）分子对接应用

分子对接应用一般基于高通量计算框架来搜寻与受体大分子具备最佳结合模式的配体小分子，配体小分子则来自多个数据中心的用户共享数据集。针对此场景，在笔者研发的虚拟数据空间系统中，分子对接应用可透明地实现多中心数据聚合能力，将所有分子数据从逻辑上聚合起来，给用户提供统一的数据视图，直接以文件系统的形式访问不同中心的数据，同时也可以利用存算联动机制将计算任务合理分发到对应数据所在的超算中心，以减少数据迁移，实现计算结果的自动规约。具体验证情况如图10所示。

该应用在验证环境中的4个节点上部署，验证结果表明，吞吐率达到了单个节点的3.07倍，有效提升了分子对接应用的执行效率。

（2）全基因组关联分析应用

全基因组关联分析需要处理大规模数据，计算过程中会使用多个计算工具，产生大量阶段性计算的中间文件。单个分析数据文件达数百兆，且与基因测序深度和测序人数相关，深度越大，人数越多，数据量越大，一般在几百TB到PB级。实验所用基因数据测序深度为0.1×时，实验中用到的平均单个基因文件约为260 MB，测序100万人的基因组就需要处理100万个基因文件，

图 8 AROSE 平台集成

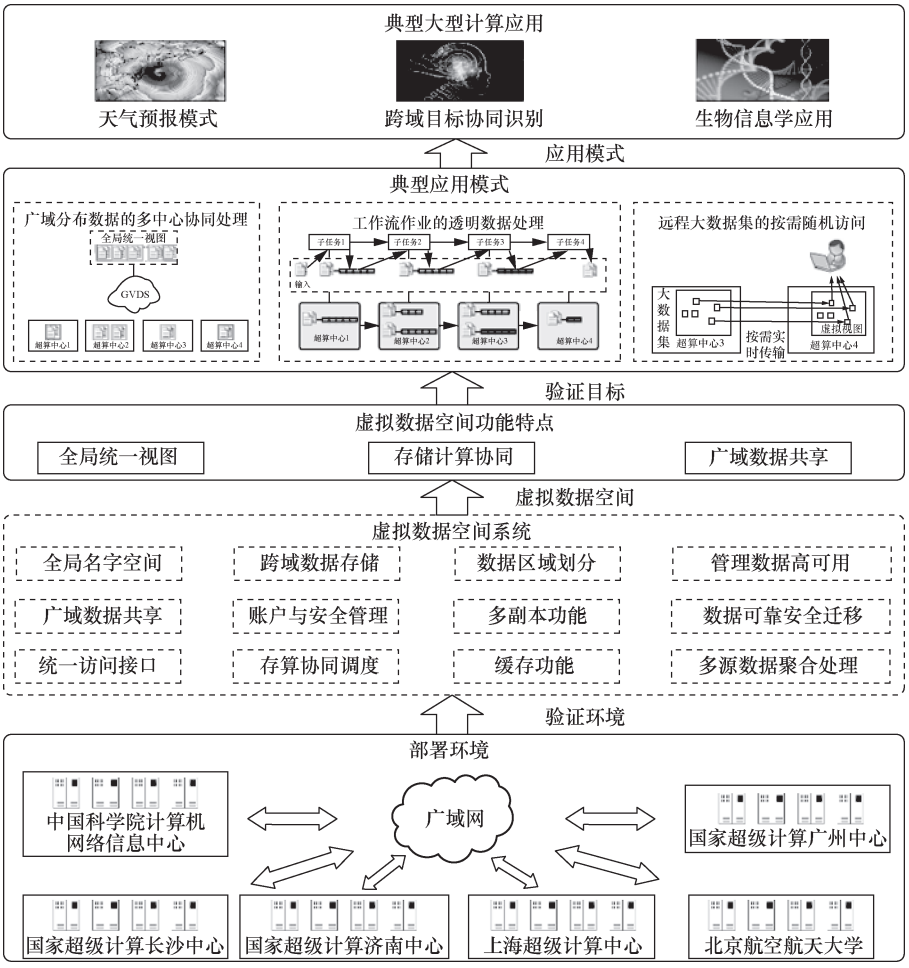


图 9 验证方案

(a) 执行过程 (b) 运行时间

图 10 分子对接应用验证

数据量达几百TB。若将测序深度提高到1×，则数据量可达PB级。此外，分析过程中用到的多个计算工具之间也不可避免地会产生中间文件，用于计算工具衔接和避免程序崩溃，从而在程序崩溃时不必重启全部计算。这就使得在原有数据规模的基础上，文件量又成倍地增加。如此大量的文件访问使得元数据服务器极易拥堵，因为在分布式文件系统中，相比对象存储服务器，元数据服务器更容易成为瓶颈。针对此场景，笔者在部署环境上进行了相关实验，具体如**图11**所示。

测试中，虚拟数据空间为超算中心的生物数据库建设提供了支撑，一方面汇聚了各超算中心用户提供的开放数据，可供更多科研工作者共享；另一方面为应用提供了跨多个超算中心提取数据的能力，并针对一些特定的数据查询、匹配操作，采用存算协同机制，将计算任务分发到多个超算中心，以提高并发性。

（3）天气预报模式应用

天气研究与预报（weather research and forecasting，WRF）模型是典型的中尺度天气预报模式和同化系统，属于计算密集型应用，数据量小，但计算量大，需持续将数据输入计算中心，而将各气象站数据以文件传输方式汇聚到计算中心是一项繁杂的工作。此外，为了更精确地预测气候变化，模式的精度和分辨率需求也在不断提高，这使得模式的计算量大幅增加。

验证中采用基于嵌套降尺度的WRF应用，在需要计算的区域嵌套多层、多块不同分辨率的网格，细网格通过相邻粗网格根据细化率进行局部加密得到，从而将中心A和中心B的计算时间重叠，缩短整体计算时间，以更好地利用各中心的闲置资源，协同完成大尺度、高分辨率的天气预报，具体如**图12**所示。

虚拟数据空间的全局虚拟视图可将多采集点数据逻辑汇聚到虚拟数据空间，各时序任务从虚拟数据空间获取数据。在交互时，从虚拟数据空间查看所需数据的生成及完整性，传统模式则通过ssh远程查询。通过虚拟数据空间访问远程数据，数据的迁移和读取可由系统自动完成，基于系统提供的存算调度能力，也可以自适应地选择数据向任务迁移，或者任务向数据迁移，以提高资源利用率，避免跨广域的通信开销。

4.4 系统性能综合测试

为了验证虚拟数据空间系统的可靠性，笔者还开展了性能综合测试，测试模

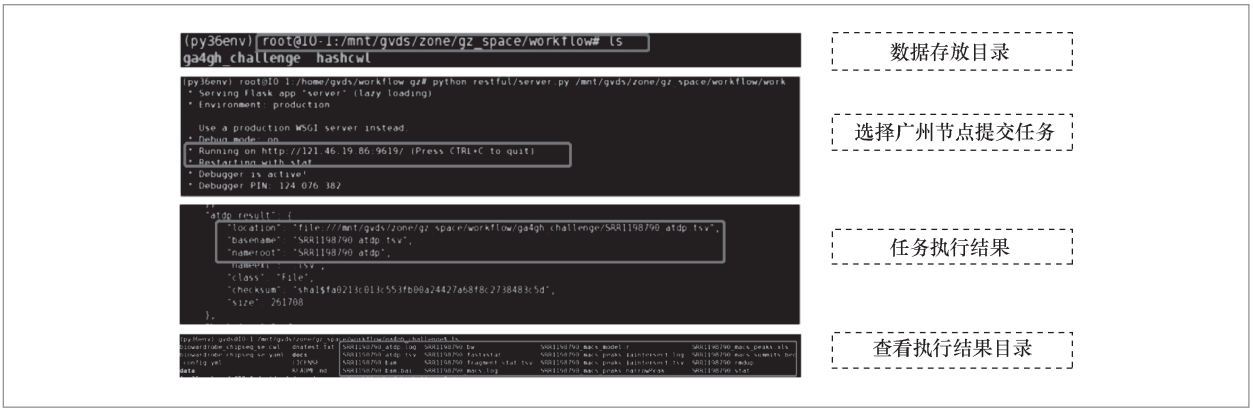


图 11 全基因组关联分析

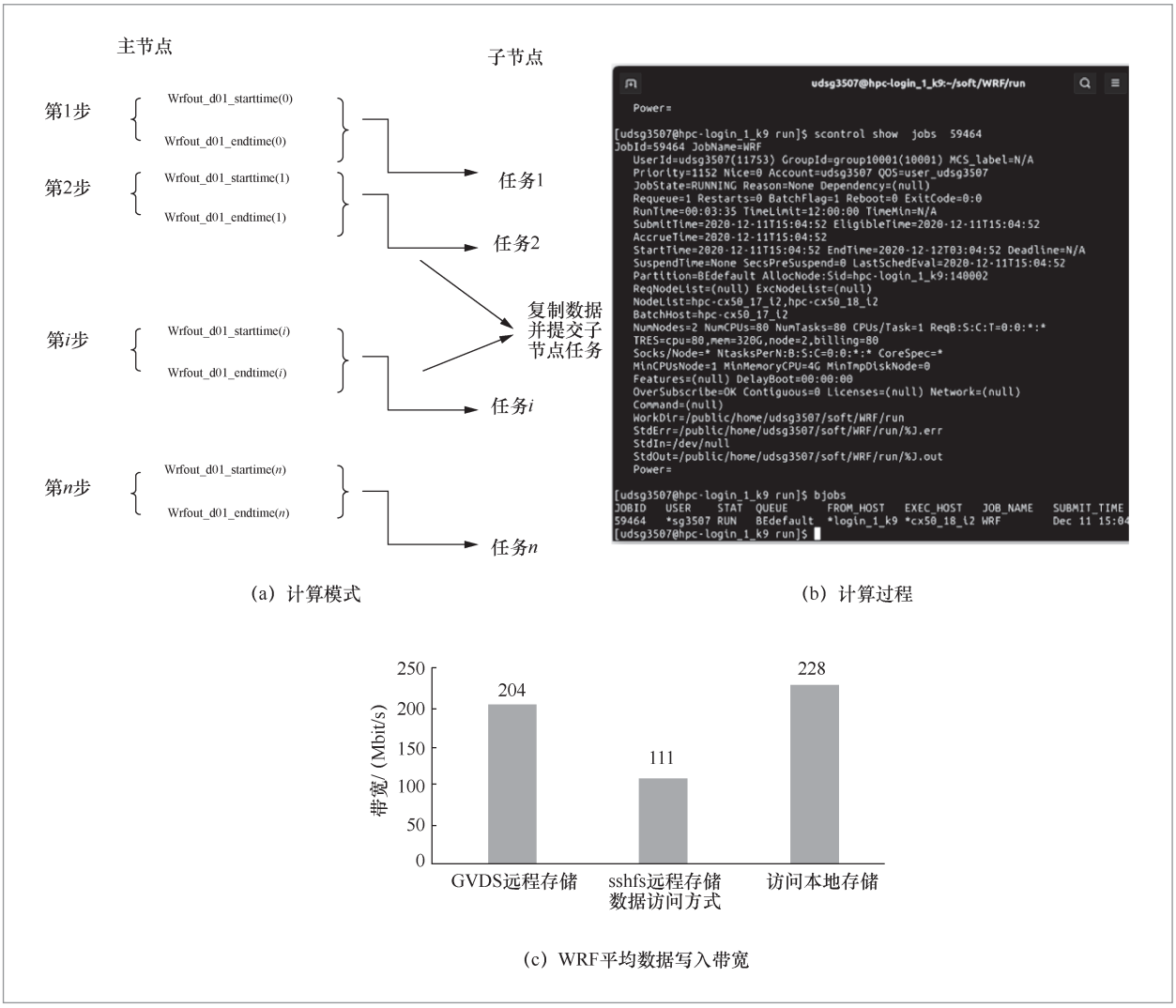


图 12 天气预报模式应用

式如**图13**所示。以国家超级计算济南中心为中心，从国家超级计算长沙中心、国家超级计算广州中心、上海超级计算中心和中国科学院计算机网络信息中心以总负载压力超过1 GB/s的多类型负载压力对国家超级计算济南中心进行了为期25天的不间断访问。

测试采用FIO、DD等压力测试工具，从广域网中不同超算中心的多个客户端产生混合负载，不间断访问远程中心，测试结果如**图14**所示。

测试中，中国科学院计算机网络信息中心到国家超级计算济南中心的吞吐量稳定在105 MB/s左右，上海超级计算中心到中国科学院计算机网络信息中心为50 MB/s左右，国家超级计算长沙中心到上海超级计算中心为35 MB/s左右，这是因为节点对间的物理带宽不一样。测得的节点对间广域网带宽如**图15**所示。

由于笔者研发的系统也提供了自适应数据缓存、按需远程访问、数据块级访问等能力，而且也对广域网通信做了大量并发通信方面的优化，测试结果基本上能保持在

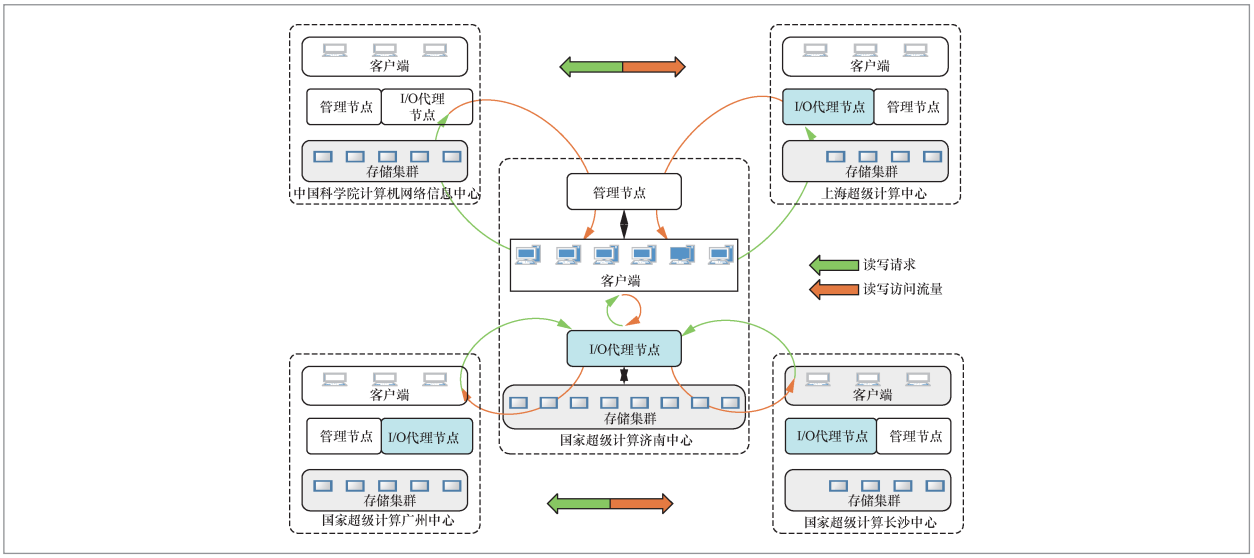


图 13 测试模式

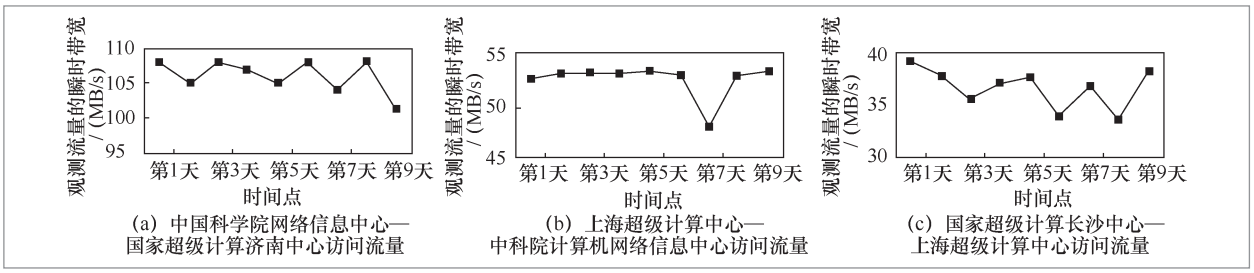


图 14 稳定性测试

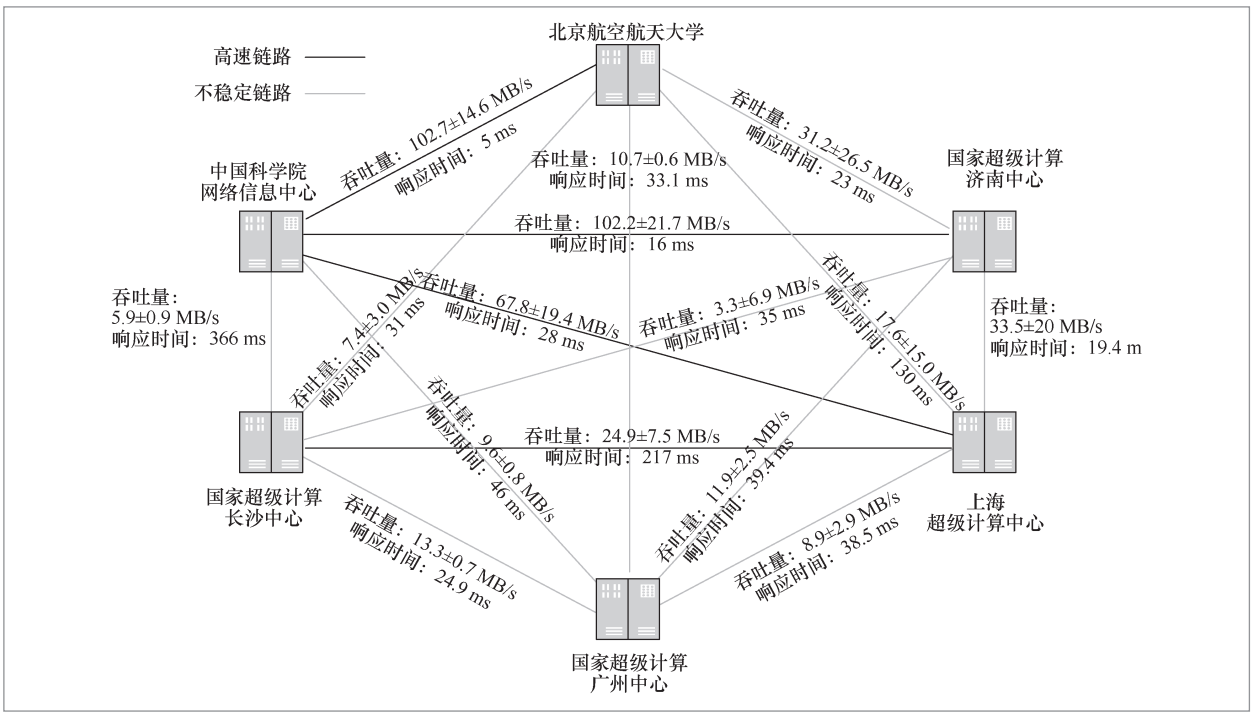


图 15 节点对间的广域网带宽

物理带宽附近。这一方面表明在高负载压力和长时间运行过程中，系统仍然能够较好地保证可靠性和吞吐量的稳定性；另一方面也表明系统在数据访问和广域网带宽优化等方面的关键技术行之有效。

5 讨论

本文设计的虚拟数据空间系统架构在国家高性能计算环境中，可统一管理和利用国家高性能计算环境计算和存储资源。目前，虚拟数据空间已部署并汇聚了国家高性能计算环境5个超算中心的存储资源，通过统一名字空间进行统一管理。此外，也可在计算时通过存算联动策略选择最佳的用户计算策略，并通过虚拟数据空间对用户的计算作业和数据进行调度，从而实现对国家高性能计算环境的资源汇聚及提升。

基于本文研究成果，预期可有效提高跨中心协同工作的效率，并推动国家高性能计算环境中大型应用跨域计算模式的发展，同时提高全系统的资源利用率。无虚拟数据空间下的数据访问方式（即当前的跨域文件访问方式）如图16所示。

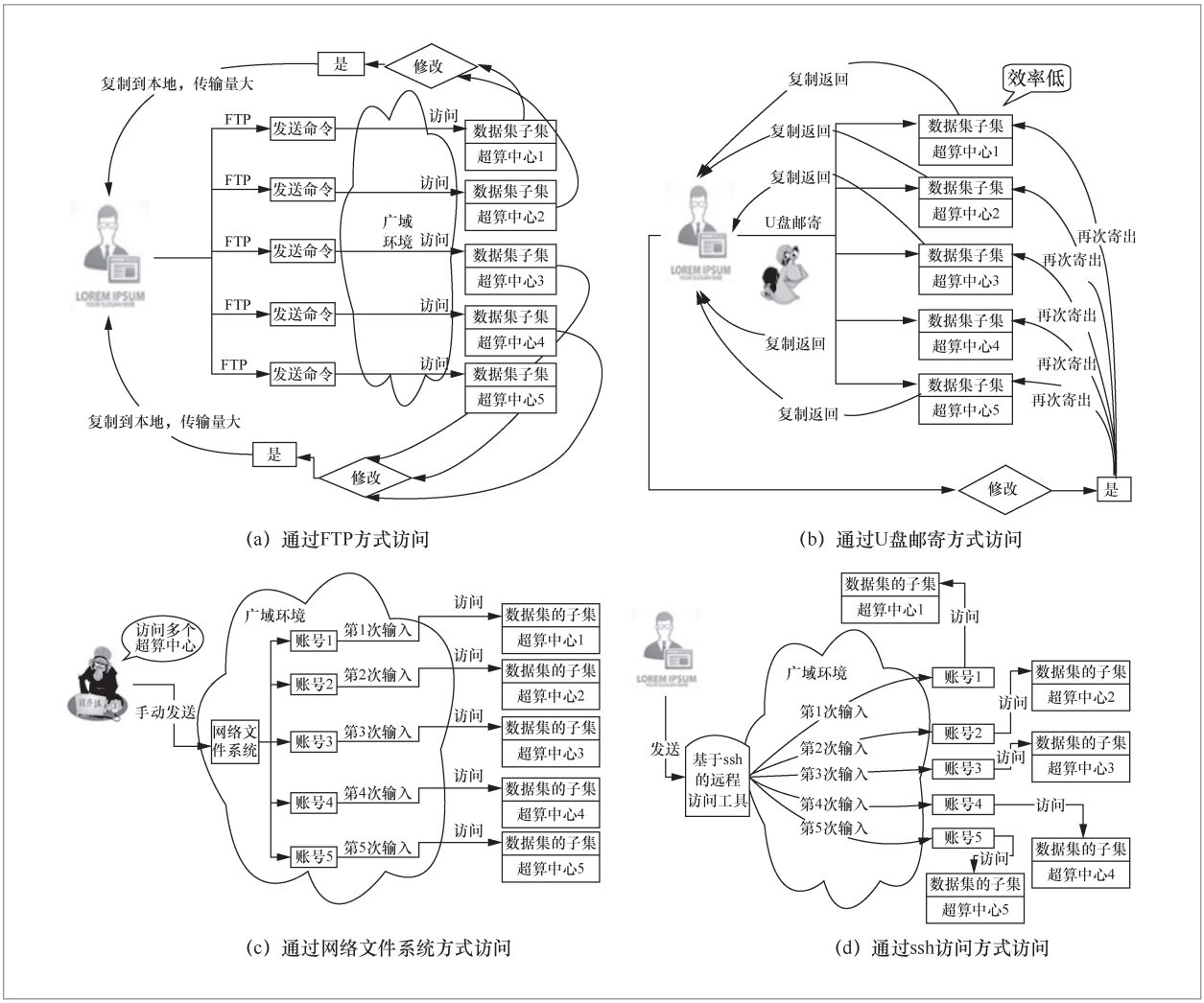


图 16 当前的跨域文件访问方式

目前对于跨广域环境的基本文件操作，一般通过FTP、U盘邮寄、网络文件系统、ssh访问等方式实现，导致数据传输量大、时延大，且需要用户手动通过多个账号与各个超算中心进行连接和登录。基于本文的虚拟数据空间，则可以通过一站式登录远程访问并执行与本地访问一致的操作，同时也可实现存储和计算的协同调度及数据和作业合理全局放置，如图17所示。

例如，在刑侦、安防等大型应用中，搜寻和追踪一个目标时往往涉及跨地域的多计算中心和多数据源，利用虚拟数据空间实现的多中心数据聚合能力，可以给用户提供统一的数据视图，并通过存储计算协同机制将计算任务合理分发到对应数据所在的中心，以减少数据迁移，实现计算结果的自动规约。

6 结束语

本文针对国家高性能计算环境中聚合广域分散存储资源的技术短板及大型计算应用对跨域全局虚拟数据空间的现实需求，建立了一套可用于构建广域全局虚拟数据空间的完整技术体系，研发了一个可运行于国家高性能计算环境的虚拟数据空间系统。该成果从核心技术层面解决了长期困扰我国高性能计算环境发展的广域存储管理访问瓶颈问题，填补了我国在广域分散存储资源统一管理和跨域访问方面的技术空白，为在国家高性能计算环境中建立跨域虚拟数据空间提供了技术手段和应用经验。对于推动完善我国自主高性能计算环境软件技术体系，支撑建设资源共享、统一管理、高效协同的国家高性能计算环境，促进我国高性能计算环境自主可控和可持续发展具有重要意义。

笔者的下一步工作是进一步提升国家高性能计算环境的部署规模和系统的功能扩展，并开展用户推广和宣传工作，推动研究成果与现有国家高性能计算环境的深度融合，高效聚合广域分散资源，充分发挥资源聚合效应，有效支撑大型计算应用，促进我国高性能计算环境及应用的可持续发展。

致谢

感谢国家重点研发计划“高性能计算虚拟数据空间”项目团队的各位老师和同学，以及为项目研发提供指导的各位项目专家。

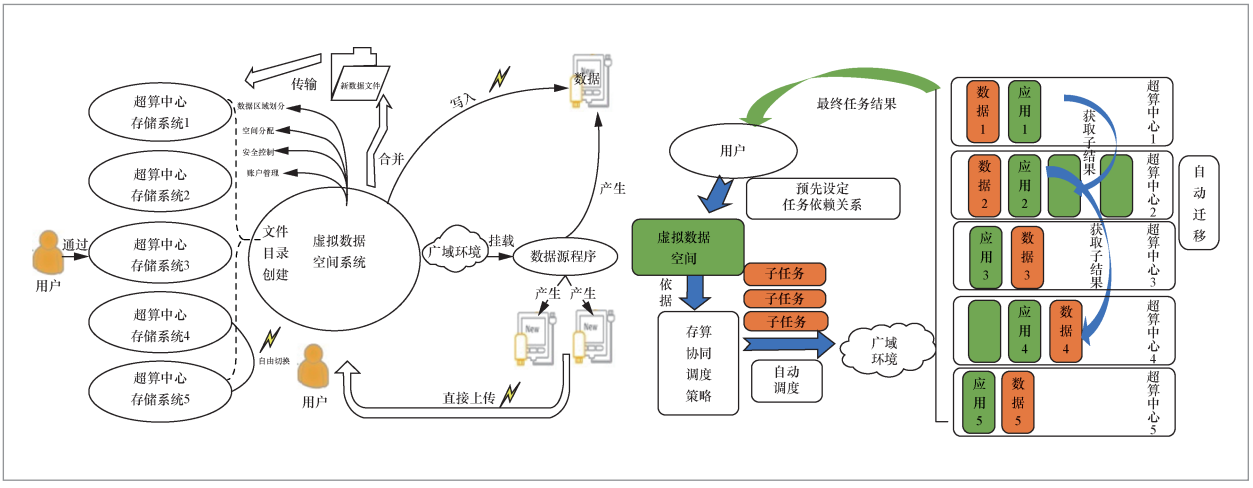


图 17 基于虚拟数据空间的应用访问与计算模式

参考文献:

[1] QIAN D P. High performance computing: a brief review and prospects[J]. National Science Review, 2016, 3(1): 16.

[2] VILJOEN M, DUTKA Ł, KRYZA B, et al. Towards European open science commons: the EGI open data platform and the EGI dataHub[J]. Procedia Computer Science, 2016, 97: 148–152.

[3] WRZESZCZ M, TRZEPLA K, SOTA R, et al. Metadata organization and management for globalization of data access with OneData[C]// International Conference on Parallel Processing and Applied Mathematics. Heidelberg: Springer, 2015: 312–321.

[4] GRIMSHAW A, MORGAN M, KALYANARAMAN A. GFFS—the XSEDE global federated file system[J]. Parallel Processing Letters, 2013, 23(2): 134005.

[5] CATLETT C, ALLCOCK W E, ANDREWS P, et al. TeraGrid: analysis of organization, system architecture, and middleware enabling new types of applications[M]// High Performance Computing and Grids in Action. Amsterdam: IOS Press, 2008.

[6] TOWN J, BOISSEAU J, ROSKIES J, et al. XSEDE: extreme science and engineering discovery environment (OAC 15–48562)[R]. 2020.

[7] NEWHOUSE S. Seeking new horizons: EGI's role in 2020 (EGI–1098–D230–V3)[R]. 2021.

[8] KUBIATOWICZ J, BINDEL D, CHEN Y, et al. OceanStore: an architecture for global–scale persistent storage[J]. ACM SIGPLAN Notices, 2002, 35(11).

[9] MAYMOUNKOV P, MAZIÈRES D. Kademlia: a peer–to–peer information system based on the XOR metric[C]// The 1st International Workshop on Peer–to–Peer Systems. Heidelberg: Springer, 2002: 53–65.

[10] CORBETT J C, DEAN J, EPSTEIN M, et al. Spanner: Google's globally–distributed database[J]. ACM Transactions on Computer Systems, 2012, 31(3): 8.

[11] THOMSON A, ABADI D J. CalvinFS: consistent WAN replication and scalable metadata management for distributed file systems[C]// The 13th USENIX Conference on File and Storage Technologies. Berkeley: USENIX Association, 2015: 1–14.

[12] WU Z, BUTKIEWICZ M, PERKINS D, et al. SPANStore: cost–effective geo–replicated storage spanning multiple cloud services[C]// The 24th ACM Symposium on Operating Systems. New York: ACM Press, 2013: 292–308.

[13] BERMBACH D, KLEMS M, TAI S, et al. MetaStorage: a federated cloud storage system to manage consistency–latency tradeoffs[C]// The 2011 IEEE International Conference on Cloud Computing. Piscataway: IEEE Press, 2011: 452–459.

[14] CALDER B, WANG J, OGUS A. Windows Azure Storage: a highly available cloud storage service with strong consistency[C]// The 23rd ACM Symposium on Operating Systems. New York: ACM Press, 2011: 143–157.

[15] HENSCHEL R, SIMMS S, HANCOCK D, et al. Demonstrating Lustre over a 100Gbps wide area network of 3500km[C]// The International Conference on High Performance Computing, Networking, Storage and Analysis. Piscataway: IEEE Press, 2012: 1–8.

[16] TATEBE O, HIRAGA K, SODA N. Gfarm grid file system[J]. New Generation Computing, 2010, 28(3): 257–275.

[17] LANDERS M, ZHANG H, TAN K. PeerStore: better performance by relaxing in peer–to–peer backup[C]// The 4th International Conference on Peer–to–

Peer Computing. Piscataway: IEEE Press, 2004: 72–79.

[18] CAO W, LIU Z J, WANG P, et al. PolarFS: an ultra–low latency and failure resilient distributed file system for shared storage cloud database[J]. Proceedings of the VLDB Endowment, 2018, 11(12): 1849–1862.

[19] 胡进锋, 洪春辉, 郑纬民. 一种面向对象的Internet存储服务系统Granary[J]. 计算机研究与发展, 2007, 44(6): 1071–1078.
HU J F, HONG C H, ZHENG W M. Granary: an architecture of object oriented Internet storage service[J]. Journal of Computer Research and Development, 2007, 44(6): 1071–1078.

[20] ZHANG Z J, XIAO L M, SU S B, et al. HSAStore: a hierarchical storage architecture for computing systems containing large–scale intermediate data[C]// International Conference on Collaborative Computing: Networking, Applications and Worksharing. Heidelberg: Springer, 2017: 591–601.

[21] XIAO H, WU H, CHI X. SCE: grid environment for scientific computing[C]// International Conference on Networks for Grid Applications. Heidelberg: Springer, 2008: 35–42.

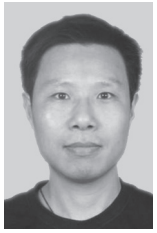
[22] 胡正丁, 薛巍. 面向异构众核超级计算机的大规模稀疏计算性能优化研究[J]. 大数据, 2020, 6(4): 40–55.
HU Z D, XUE W. Research on performance optimization for large–scale sparse computation over many–core heterogenous supercomputer[J]. Big Data Research, 2020, 6(4): 40–55.

[23] 韦冰. 面向广域高性能计算环境的文件数据访问和容错方法研究[D]. 北京: 北京航空航天大学, 2020.
WEI B. A study on file data access and fault tolerance in the wide area high performance computing environment[D]. Beijing: Beihang University, 2020.

[24] 周汉杰. 广域虚拟数据空间副本技术研究与实现[D]. 北京: 北京航空航天大学, 2020.
ZHOU H J. Research and implementation of replication technology for global virtual data space[D]. Beijing: Beihang University, 2020.

[25] SONG Y, XIAO L M, WANG L, et al. GCSS: a global collaborative scheduling strategy for wide–area high–performance computing[J]. Frontiers of Computer Science, 2021, accepted.

作者简介



秦广军(1977–), 男, 博士, 北京联合大学智慧城市学院讲师, 中国计算机学会会员, 主要研究方向为高性能计算、存储系统、大数据和机器学习等。作为项目骨干参与了国家863计划项目、国家重点研发计划项目、国家自然科学基金项目、北京市自然科学基金项目等。

肖利民(1970–), 男, 博士, 北京航空航天大学计算机学院教授、博士生导师, 计算机科学技术系主任, 计算机系统结构研究所副所长, 中国计算机学会大数据专家委员会委员、高性能计算专业委员会常务委员、容错计算专业委员会委员, 中国电子学会云计算专家委员会委员, 国家计算机科学技术名词审定委员会委员, 国家科技基础条件平台专家组成员, 工业和信息化部电子科学技术委员会委员, 中国工程院中国信息与电子工程科技发展战略研究中心专家委员会特聘专家。主要研究方向为计算机体系结构、计算机软件系统、高性能计算、云计算、虚拟化技术等。先后获得国家科技进步奖二等奖、北京市科学技术奖一等奖、中国科学院科技进步奖一等奖、原信息产业部信息产业重大技术发明奖、科技部国家重点新产品奖等国家级和省部级科技奖励。

张广艳（1976– ），男，博士，清华大学计算机系长聘副教授、博士生导师，主要从事大数据存储与分析的理论和方法研究，包括大数据计算、存储系统与分布式处理等方面。研究得到了国家杰出青年科学基金项目、国家重点研发计划项目、国家973项目和国家863项目等的支持。近年来提出了大规模存储系统构建及访问的方法与关键技术，有效提高了存储系统的性能、扩展性和可用性。发表学术论文40余篇，其中在FAST、USENIX ATC、*ACM TOS*、*IEEE TC*、*IEEE TPDS*等计算机系统领域高水平国际会议和期刊发表论文20余篇。近五年以第一发明人获得美国发明专利授权1项、中国发明专利授权7项。

牛北方（1978– ），男，博士，中国科学院计算机网络信息中心研究员，中国科学院大学岗位教授、博士生导师。中国计算机学会高性能计算专业委员会委员。主要研究方向为高性能计算、数据分析算法与软件技术。

陈志广（1984– ），男，博士，中山大学计算机学院副教授，主要研究方向为大数据存储与处理、并行与分布式计算、高性能计算与超级计算机。

收稿日期： 2021-01-20
通信作者： 肖利民，xiaolm@buaa.edu.cn
基金项目： 国家重点研发计划资助项目（No.2018YFB0203901）
Foundation Item: The National Key Research and Development Program of China (No.2018YFB0203901)