高能物理实验的离线计算

李卫东　石京燕　汪　璐　张晓梅　程耀东　齐法制　曾　珊　颜　田

(中国科学院高能物理研究所　100049)

高能物理研究组成物质的基本粒子及其相互作用规律，是物理学研究中的最前沿。当今，高能物理实验规模一般都很大，需要成百上千的科学家参加。高能物理实验的周期比较长，从实验设计到目标的实现通常会经历十几年甚至几十年的时间。实验产生的海量实验数据，需要借助先进的计算机技术来处理和分析，实验的需求也助推了计算机信息技术的不断发展。近年来，我国物理学家在以我为主的高能物理实验中取得了令人瞩目的成绩，其中包括北京正负电子对撞机实验和大亚湾反应堆中微子实验。下面我们将以这两个实验为例，介绍数据存储、数据传输以及各种计算技术在高能物理实验中的运用。

一、数据处理与分析

通过触发判选和在线选择的事例，由在线数据获取系统以二进制文件的形式记录下来。这种数据称作原始数据，主要包含探测器电子学信号的时间和幅度信息。通过高速以太网，原始数据文件被传输到磁带库永久保存。对原始数据进行刻度和重建后，生成重建数据，供物理分析使用。

离线数据处理和物理分析的简化过程如图 1 所示。原始数据经过离线刻度，能够消除实验的各种外部条件（例如温度、气压）和探测器本身条件（例如探测器高压）对电子学信号与物理测量量之间转换关系的影响。离线刻度将按不同的子探测器分别进行，生成的大量刻度常数保存于数据库。重建是离线数据处理的核心，数据重建算法使用刻度算法产生的刻度常数，将探测器记录的原始数据转化为粒子的动量、能量和运动方向等物理量，生成重建数据。物理研究还需要产生与真实数据数量相当的模拟数据，这部分数据也要进行重建。和原始数据一样，所有重建数据会被保存在磁带库中。物理分析人员利用物理分析工具例如运动学拟合、粒子衰变顶点寻找和粒子鉴别等软件，分析重建数据，得到物理研究结果。

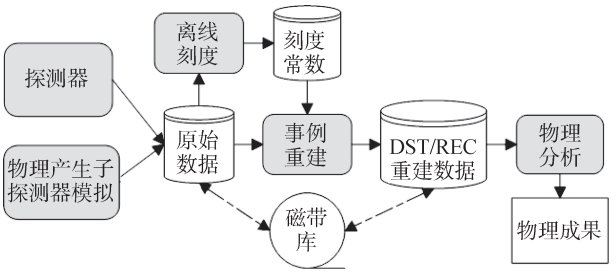


图 1　离线数据处理流程

二、数据存储与传输

高能物理计算属于数据密集型高性能计算，数据存储系统是影响计算性能的关键环节。数据存储系统不仅要保存海量数据，同时还要考虑与数据处理系统的配合，提高数据分析效率。大部分高能物理计算是高吞吐率的计算 (High Throughput Computing, HTC)，追求系统整体而非单个作业的性能和效率。这里吞吐率指一个计算机或数据处理系统单位时间内的数据处理量或传输量。在表示数据量的大小时，常用的单位有 kB(10^3 Bytes), MB(10^6 Bytes), GB(10^9 Bytes) 和 TB(10^{12} Bytes)。 在表示特别大的数据量时，还会用到 PB(10^{15} Bytes) 和 EB(10^{18}Bytes)。高能物理数据分析的读写（Input/Output, 简称 I/O）模式以大文件（数百 MB 甚至 GB 级）、大块 (MB 级记录块) 读写、一次写多次读、吞吐率需求高（单个作业需要几 MB/s）为特征。同时，物理学家对大量小文件（kB 级的程序和文档）的查找和浏览也对元数据访问性能提出了很高的要求。

高能物理数据以非结构化数据为主。目前，常用的非结构化数据存储系统包括集群文件系统、应用

层存储系统和分级存储系统等。这三者都采用了分布式存储技术，本身并没有非常严格的区分，只是关注的侧重点有所不同。集群文件系统一般以传统文件系统的方式来访问，客户端实现内核模块，完全兼容 POSIX 语义，因此上层的数据处理软件无需任何修改即可使用海量的存储空间，能够很好地兼容原有应用。常见的集群文件系统包括 Lustre、Gluster、GPFS、ISILON 等，其中全世界最快的超级计算机（TOP500）中有 70% 以上都在使用 Lustre 系统。应用层存储系统一般不实现文件系统内核模块，不完全兼容 POSIX 语义，针对特定的应用场景进行优化，因此往往表现出更好的可扩展性和性能，但是上层应用程序必须要调用特定的应用程序接口（API）才能访问。分级存储系统是指根据文件的访问频率、热度等因素，将不同的文件分配到不同的存储设备上存放。基于磁盘 – 磁带的分级存储系统比较成熟，比如 CASTOR、dCache 等系统广泛应用于高能物理领域。当前，基于固态硬盘 (SSD) 和串口机械硬盘（SATA）做分级存储是研究热点，如开源项目 flashcache 和扩展项目 flashcachegroup 等。现有的分布式存储系统还有谷歌文件系统（Google File System，GFS）和分布式文件系统（Hadoop Distributed File System，HDFS）等，其中 HDFS 是一套开源软件，在互联网的大数据存储中应用尤为广泛。科研大数据的存储量往往达到 PB 级甚至更高，因此存储的成本和性价比也是重要的考虑因素。为了使用部分云计算资源以及解决数据的异地复制需求，高能物理计算领域也在考虑云存储技术与计算框架的结合和性能优化。

高能物理研究所（以下简称高能所）的计算环境中，存储系统分为磁盘文件系统 Lustre 和分级存储系统 Castor 两个部分，如图 2 所示。数千个计算节点和近百个存储服务器之间通过万兆以太网络连接，存储软件为计算作业屏蔽了复杂的后端架构，用户可以像使用单机存储设备一样使用海量存储空间。目前，右侧的 Lustre 磁盘存储系统包括 50 多台数据服务器，100 多台磁盘存储阵列，能够提供约 3 PB 存储空间，40 GB/s 的峰值聚合带宽。同时，计算中心开发了自动优化、进程快照、行为分析、故障报警等附加功能，提高系统的自适应性、可靠性和管理效率。左侧的分级存储系统 Castor 用于存放不频繁访问、需要长期保存的数据，例如备份数据，原始物理数据等。系统采用 IBM Total Storage 3584 智能磁带库和 LTO4 磁带，可存放 6000 多盘磁带，提供 5 PB 以上的存储空间。目前，系统能够提供 90MB/s 单驱动器读写性能，2GB/s 的聚合读写性能。

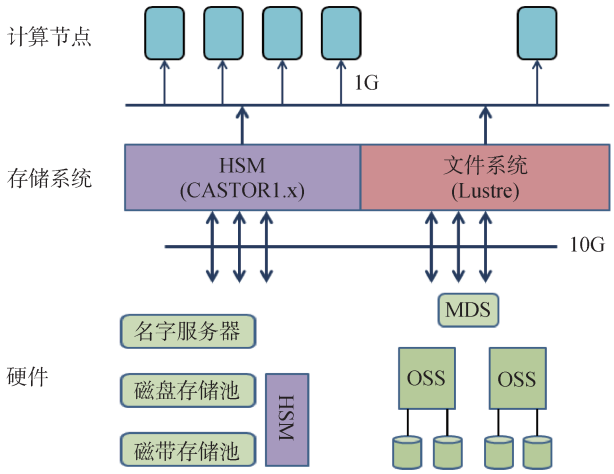


图 2　高能所的存储系统架构

在实际应用中，单个存储设备很难满足高能物理计算 PB 甚至 EB 级的存储和数十 GB/s 乃至 TB/s 的吞吐率需求，高能物理数据存储系统必须是分布式、多服务器、多设备的。在一个庞大的网络连接的系统中，设备故障、网络中断和延时、服务器死机是常态。因此高能物理计算对存储系统的可扩展性、易用性、数据可靠性和高可用性提出了不小的挑战。同时，考虑到存储需求的递增性和存储设备的更新换代，存储资源总是逐步扩张的。存储系统软件还必须很好地解决性能的可扩展性以及数据的自动负载均衡问题。

高能物理实验每天都会产生大量的实验数据，部分高能物理实验本身具有跨地域建设特性，这些实验数据需要传输到远程的数据和计算中心进行离线分析，如何将这些数据实时、可靠、高效地传输到远程的数据和计算中心则是目前高能物理实验中需要解决的一个重要问题。

目前高能物理数据传输系统大多数都基于支持并发传输的工具（如 GridFTP、bbftp 等）来实现，其基本框架如图 3 所示，以大亚湾数据传输系统为例，现

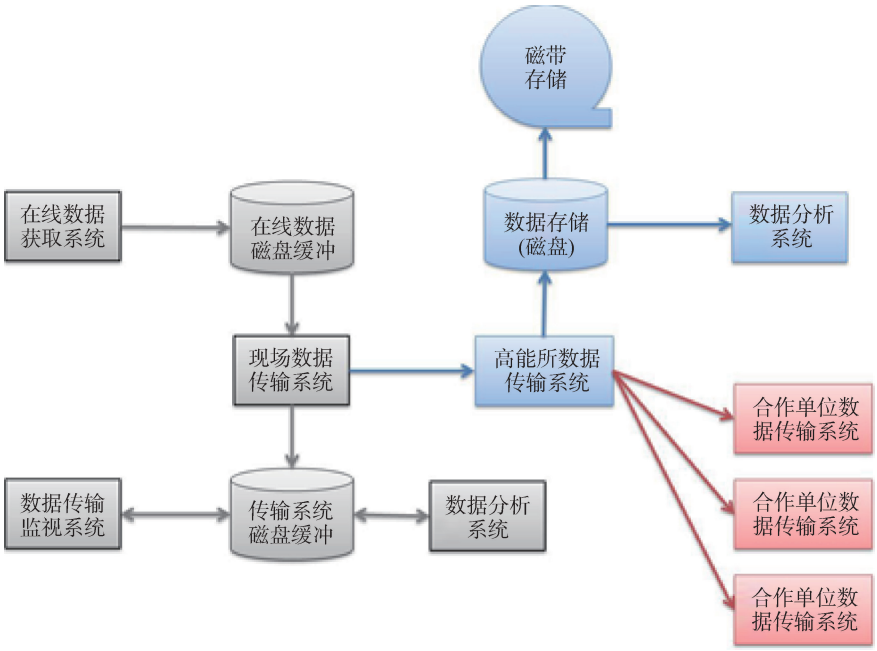


图 3　数据传输系统部署架构图

场的数据传输系统将在线数据获取系统中的数据远程传输到高能所计算中心，并保存在分布式并行文件系统和数据备份系统中，然后再将数据分发到其他合作单位，以便全球的科学家进行数据分析和处理。

为了保证数据传输的可靠性，数据传输系统都具有传输过程管理和传输性能监控的功能。数据传输系统提供图形化的监视模块对数据传输量、传输效率和可靠性等参数进行实时监控和分析，如图 4 所示。

为了保证数据交换的高效性，数据传输系统的性能也依赖于传输链路上的广域网性能，目前，高能所已经和各合作组成员国之间建立了良好的广域网链路，是国际网络出口带宽最大的研究机构，如图 5 所示：大亚湾、羊八井、东莞采用专线将数据传输到北京，带宽为 155 Mbps；高能所经过伦敦到欧洲共享带宽为 5 Gbps，经过清华大学到美国的共享带宽为 10 Gbps。

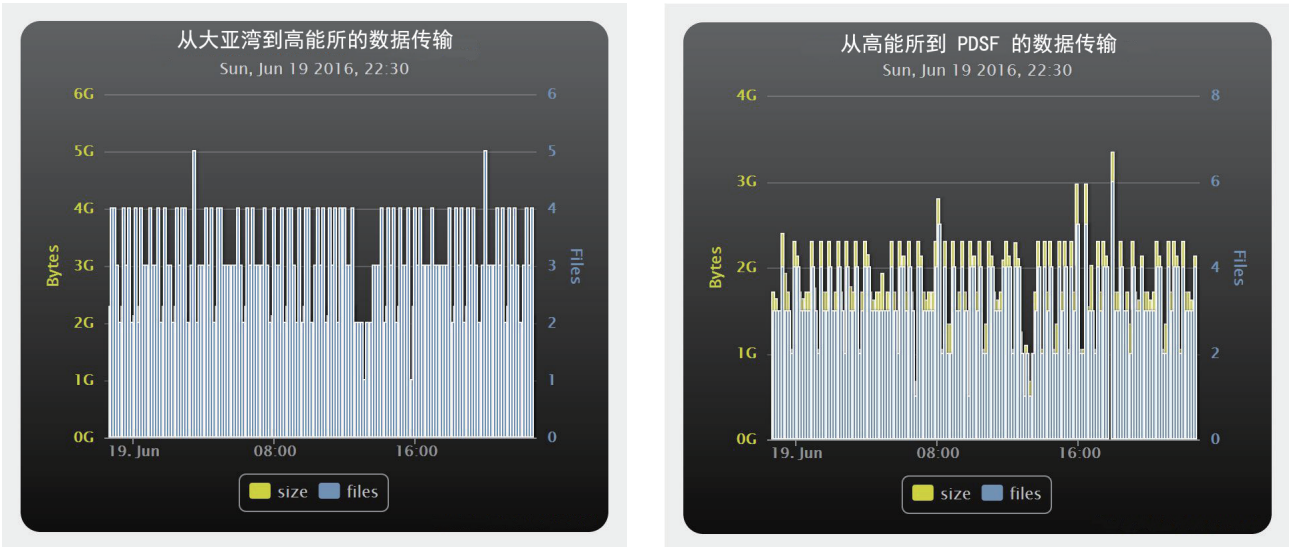


图 4　数据文件传输过程效果监控图

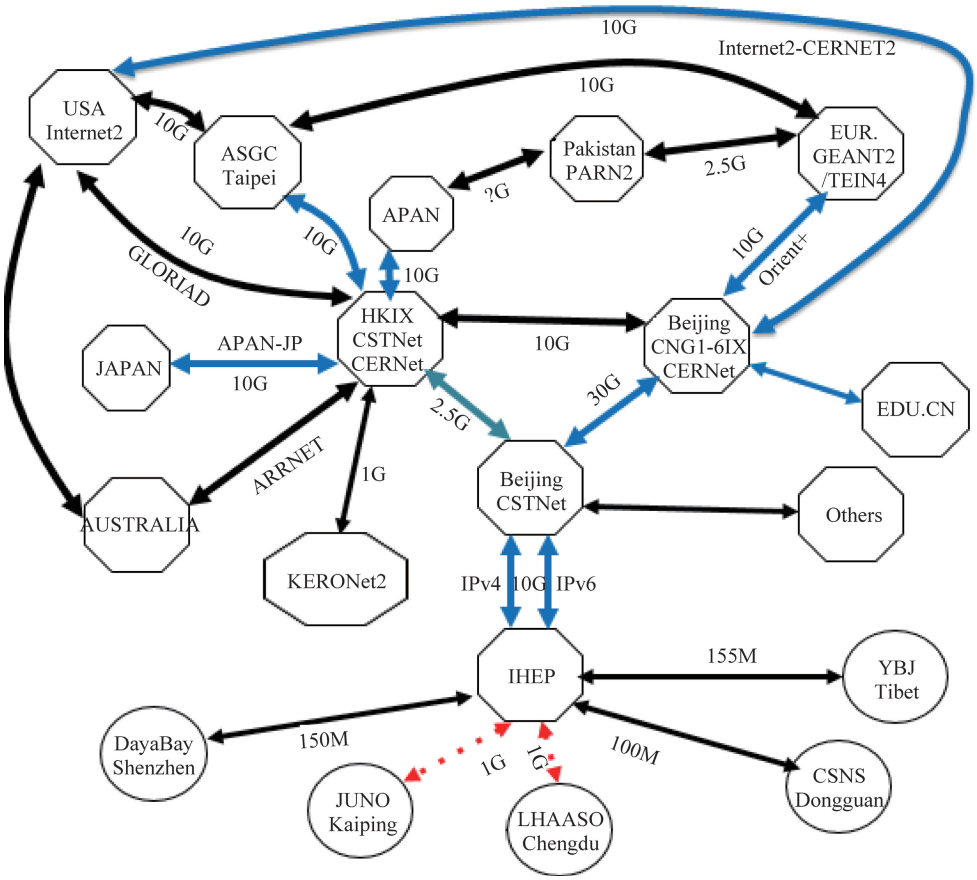


图 5　高能所广域网链路拓扑图

高能物理数据交换与共享的需求，推动着信息技术的发展，高能所于 1986 年建成中国第一条国际计算机通讯线路，并向国外发出中国第一封 Email；1988 年成为中国在国际互联网上的第一个节点；1993 年建成中国第一根国际互联网专线；1994 年建立中国第一个 WWW 网站。近几年来，高能所跟踪网络技术和架构的发展，将最新的网络技术（例如 SDN 技术、网络性能测量技术以及 40G/100G 以太网技术等）同高能物理应用需求相结合，服务于高能物理数据共享，不断提升数据交换和共享的效率。

三、数据密集型计算

高能物理实验的计算与存储需求量巨大，是典型的数据密集型计算，利用计算集群进行数据处理是高能物理计算的主要手段。计算集群是指把一组计算机通过高速网络连接在一起，构成一个整体，提供用户计算服务。一个计算集群通常由用户交互结点、计算结点、存储文件系统和资源管理作业调度服务构成。为了保证集群健壮运行，集群一般还配备有软件安装部署服务、运行监视服务和数据备份服务等。

高能物理计算是在大量物理事例中寻找极少量具有特定物理意义的事例，物理事例之间相互独立，没有相关性。通用的做法是将一批物理事例按专用的数据格式存储于数据文件中；大量高能物理数据文件由集群文件系统统一管理，提供交互结点及计算结点的读写访问。由于事例相互之间的无关性，多个不同文件可以分别被多台计算节点同时处理，计算节点之间无需相互通信，因此除了计算存储设备的硬件性能以外，计算结点数量多少也会直接影响整体数据处理速度。

一个典型的高能物理计算集群架构如图 6 所示。通过高速、可靠的网络将交互结点，计算结点，存储设备和管理服务器连接起来。按照功能不同，每个组件的软件及配置各不相同，其功能也相互独立，但整

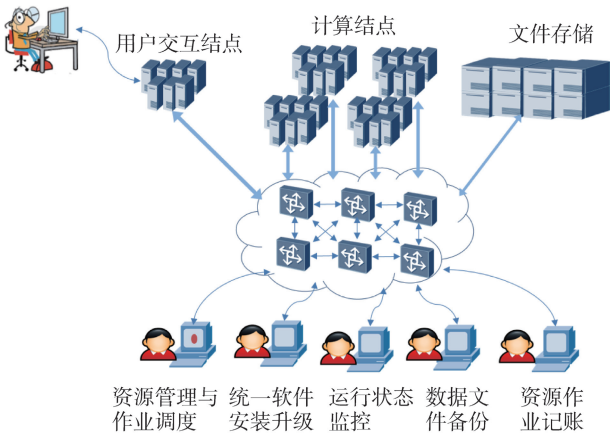


图 6　典型的高能物理计算集群

体上协同工作，提供多用户批作业计算服务。

用户在交互结点上设置各自的计算环境，编写调试程序，进行少量计算以确认程序的正确性，再将程序包装为作业后提交给计算集群。集群作业中不仅包含了需运行的程序，还有运行该程序所必需的软硬件资源需求说明。资源管理与作业调度服务是计算集群最核心的组件，它根据集群中所有计算结点的当前状态和等待运行作业的实际需求，为作业分配一个最适合的计算结点运行，此过程称之为作业调度。一个计算集群同时为很多用户提供计算服务，不同用户作业运行需求各不相同，资源管理与作业调度服务按照一定的调度策略实现作业调度。计算集群一般还需配备软件安装升级，运行监控和数据备份等管理服务器。

有些高能物理集群用 LSF、SGE 等著名的商业软件进行作业管理，除此之外一些开源的批作业调度软件由于免费易用，方便灵活等特点在高能物理领域中也得到广泛应用，其中以 Torque Maui、HTCondor、SLURM 最为有名。

Torque Maui 由最初的 PBS 批作业管理软件发展而来，曾被大量用于在高能物理计算集群。Torque 用于计算资源和作业队列管理；Maui 实现作业调度，可以提供作业回填，用户优先级等多种调度算法。但近年来此款开源软件缺少更新，用户社区不够活跃，对于大规模集群的作业调度性能不高，正在逐渐淡出使用。

HTCondor 是由美国威斯康星大学开发的一款高通量作业调度软件，它精减了复杂的调度算法，追求高效的调度性能。HTCondor 提出了分类广告板（ClassAd）机制，用于高效地匹配资源请求者（作业）与资源提供者（机器）之间需求。作业和计算节点遵循 ClassAd 机制可以非常灵活地描述各自需求与拥有属性，并由 ClassAd 进行匹配以实现作业调度。由于这种高效的调度机制非常适合高能物理计算作业简单大量的特点，被越来越多的高能物理集群所采用。

SLURM 是近年来非常活跃的一款开源软件，世界最快的大型计算机天河 II 也用其作为资源管理与调度软件。它的高度可伸缩及容错性的特点很适用大型计算集群作业调度。SLURM 以一种排他或非排他的方式为作业分配使用计算节点（取决于资源的需求）；提供框架结构启动、执行和监视作业；通过管理一个待处理工作的队列实现作业与资源管理。与 HTCondor 相比，SLURM 不仅可以支持大型计算集群的作业管理，还对 MPI 这种 CPU 密集型计算作业有着良好的支持，因此被更多科学研究计算领域采用。

四、网格计算

随着高能物理实验大数据时代的来临，原来单一的数据中心已经远远不能满足高能物理实验的数据处理和分析的计算和存储需求，高能物理对计算环境提出更高的要求：超强的计算能力和海量的数据存储能力。为了适应这一需要，一种全新的计算技术——网格计算孕育而生。互联网为高能物理实验实现了实验数据的高速共享，WWW 服务为高能物理学家实现了科研信息的充分共享，网格则是基于互联网为高能物理实验带来了计算资源和存储资源的全球共享。网格计算技术将分布在互联网上的计算资源和存储资源融合成一个整体，使得高能物理研究人员在世界上任何一个角落可以通过互联网透明地使用分布在世界上各个地方的资源，所以我们可以将网格系统比喻成一个位于全球范围的超大型计算机，如图 7 所示。

一个完整的网格系统包括安全服务、网格基础软件和网格应用软件这三个部分组成。安全服务就像网格的“卫士”，负责对进入网格系统的用户进行身份确认和访问权限确定。因此安全服务包括身份认证和权限管理两部分，其中身份认证是通过电子网格证书

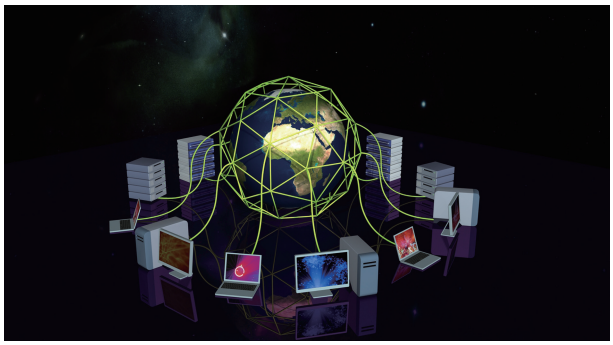


图 7　网格示意图

来实现，用户通过合法的证书签发机构（Certificate Authority，CA）申请和获得证书。位于高能所的 IHEPCA 就是由国际网格信任联盟 IGTF 认证的中国最早的 CA。网格用户是通过虚拟组织（VO）进行分组，每个实验通过虚拟组织管理系统（VOMS）对本实验用户进行管理。网格基础软件也叫网格中间件（Grid Middleware），是网格的核心部件，它建造了网格的“基础设施”，正是它实现了计算和存储资源的互联，并为网格用户提供了使用网格的基本服务，包括资源信息管理、作业管理、数据管理、监控统计等。每个加入网格系统的资源都需要安装网格中间件以保证资源被纳入统一管理和调度。得到授权的网格用户通过资源信息管理服务可以查询到可用的资源，通过作业管理服务可以进行作业的提交、查询和取回结果，通过数据管理服务可以进行数据存储、查询和获取，通过监控统计服务获取资源的状态以及使用信息。也就是说，用户可以通过统一的接口和服务，无缝地使用到网格的计算和存储资源。现在常用的网格中间件有 Globus、gLite、OSG、GOS 等几种。网格应用软件则是基于网格中间件面向特定应用和方便物理用户进行开发的软件，典型的包括大规模作业提交、实验数据集管理、实验作业监控和统计，它为最终的物理用户提供直接和专门的“服务设施”。整个网格系统的层次结构如图 8 所示。

国际上应用最广的高能物理网格平台有欧盟的 EGEE（Enabling Grids for E-sciencE）、美国的 OSG(Open Science Grid) 等。中国国家网格（CNGrid）是中国为科学实验用户提供的大型网格计算和应用平台。欧洲粒子物理中心（CERN）是最大也是最为成

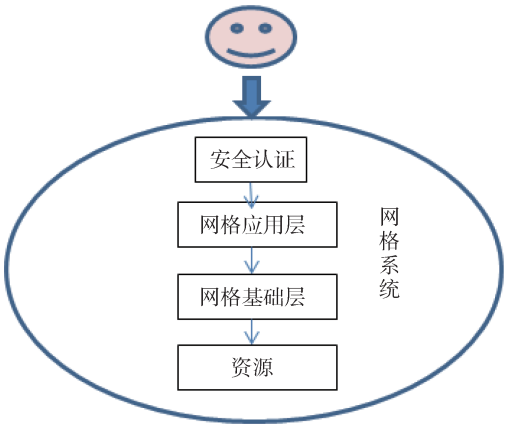


图 8　网格系统示意图

功的网格用户，基于大型强子对撞机 LHC 实验建设的 WLCG（WorldWide LHC Computing Grid）网格应用系统，包含了 42 个国家的 170 个数据中心的资源，每年处理和分享 30PB 的数据，使用了包括 EGEE 和 OSG 在内的多个网格平台，位于高能所的北京站点也是其中的一部分。WLCG 为重大物理成果——Higgs 粒子的最终发现作出了巨大的贡献。

五、云计算

云计算是一种新兴的共享基础架构的方法，近几年在产业界和学术界引起了广泛的关注。云计算是一种以服务为特征的计算模式，它通过对所有资源进行整合、抽象后以新的业务模式提供高性能、低成本的持续计算、存储及各种软件服务，支撑各类信息化应用。云计算具有资源池化、弹性可伸缩、按需自助服务、服务可计量等特征，同时具有灵活性、可靠性、可扩展性、数据集中存储、部署周期短、成本低等优势。

高能物理一直是计算技术强有力的推动者，在国际互联网、WWW 技术、网格计算的发展中都作出了积极的贡献。在云计算时代，高能物理仍然有着强烈的需求。欧洲核子中心 CERN 启动了虚拟机项目 CernVM，并在此基础上发起 LHC 云计算项目，为大型强子对撞机 LHC 提供虚拟化的应用环境。CERN 还启动了 lxcloud 项目，支持批处理计算服务，以提高资源利用率并简化管理。目前 CERN 使用 OpenStack 管理了 12 万颗 CPU 核和 1.5 万个虚拟机。德国

DESY、美国 Fermilab 等大部分国际高能物理实验室都在使用云计算技术。下面简单介绍两个典型的高能物理云计算项目：CernVM 和虚拟集群。

CernVM　2008 年，欧洲核子中心 CERN 启动了 CernVM 项目，用于解决大型强子对撞机（LHC）物理计算中的虚拟机管理问题。CernVM 的基本思想是将操作系统与应用程序打包，做成轻量级的虚拟机映像文件，从而实现在全球网格系统上的调度或是用户桌面级的数据分析。CernVM 并不是将所有的应用程序与依赖库文件都打包在一起（通常是 10GB 量级），而是初始装入大概 100MB 左右的“瘦应用”，与应用相关的程序以及数据通过 CVMFS（CernVM 文件系统）从远程软件仓库按需下载、更新和缓存，通常情况下一个应用保持在 1GB 以下。图 9 是 CernVM 的示意图。

CernVM 不仅解决了虚拟机映像文件尺寸与更新的问题，而且最大程度的保持了用户的使用习惯。CernVM 支持 VMWare、VirtualBox、Xen、KVM 等大部分主流虚拟机，可以运行在 Windows、Linux 或者 MacOS 等操作系统上。

虚拟集群　随着计算系统规模的不断扩大，操作系统与应用软件的不断升级，CPU 等硬件性能的持续提升，传统的集群或者网格计算模式面临着资源利用率不高、应用迁移复杂、多应用支持困难等问题。为此，高能所启动了虚拟集群项目。虚拟集群的系统架构如图 10 所示。底层是基于 OpenStack 的私有云。OpenStack 是一个开源的云计算管理平台，它能管理一组物理机节点上运行的虚拟机构成的资源池。这些虚拟机可以从不同的镜像启动。不同的镜像里有不同的操作系统或应用软件配置。用户可以根据需要选择合适的镜像来启动虚拟机。中间层是虚拟资源调度器，它根据任务队列情况和调度策略，弹性启动或者终止虚拟的计算节点（OpenStack 上的虚拟机）。当有新作业时，选择合适的镜像启动虚拟机；当作业完成后，关闭虚拟机，释放资源。最上层是虚拟集群队列，它将底层的云计算封装成用户熟悉的批处理队列界面，使得整个系统对用户以及基于 WLCG 的网格应用都是透明的。在用户看来，仍然是传统集群的使用方法，不必改变以前的使用方式。系统也可以支持 WLCG 网格计算等传统的高能物理计算模式。

六、结束语

高能物理实验的离线计算效率直接决定了高能物理实验物理结果的产出速度和科学发现的进程，而先进的计算机技术无疑是离线计算的“推进器”。本文介绍了高能物理实验从数据采集、存储、传输、处理和分析、最终获得物理结果的整个过程，以及前沿计算机技术在高能物理实验数据的生命周期中所起的重要作用。高能物理实验的离线计算具有数据量和吞吐量大的突出特点，先进的存储、网络和集群技术已经成为离线计算不可或缺的基本保障。我们可以看到，PB 级的并行文件系统技术已经成为海量高能物理实验数据存储和获取的必要手段，高速的万兆网络更是在连接计算资源和数据资源、实现数据在全球高能物理实验参与单位中共享的不可缺少的基础设施，集群

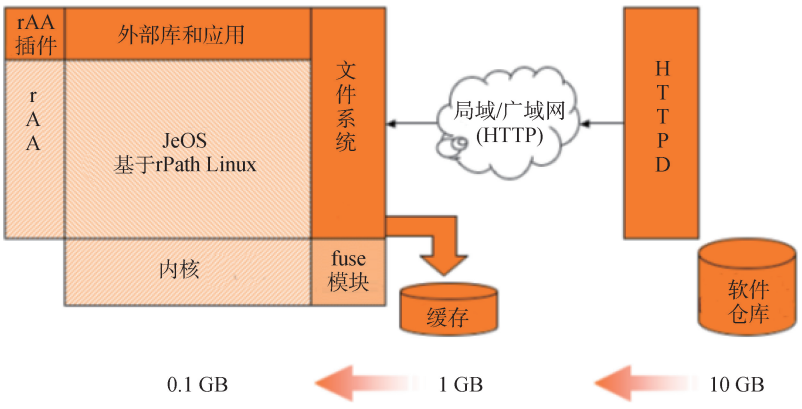


图 9　CernVM 示意图

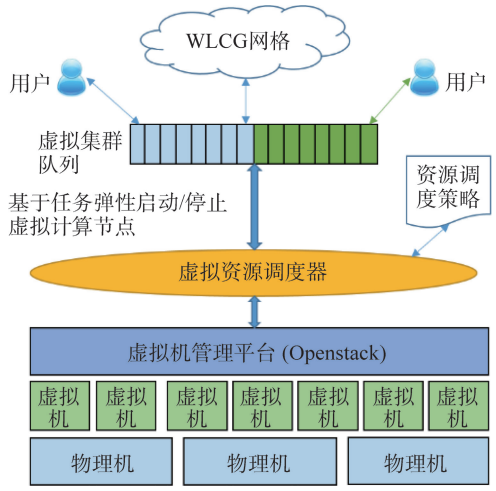


图 10　虚拟集群示意图

技术将松散的计算资源集成获得的强大的计算能力是高能物理实验数据处理与分析的必要保障。

另一方面，高能物理实验也不断推动着计算技术的创新和发展。二十年多前，高能物理实验的需求铸就了 WWW 服务的诞生。今天随着高能物理实验的规模不断扩大，数据量急速膨胀，对计算技术也提出了新的、更高的需求。现代的高能物理实验数据已经迈向 EB 量级的时代，存储和网络技术也因此需要向更快和更灵活的方向发展，出现了 EB 级存储技术、分布式存储、百万兆网络通信、网络虚拟化技术 SDN 等。同时单一的集群技术已经不能满足所有的计算需求，网格计算是又一个继 WWW 服务之后的技术变革，它使得遍布于全球的高能物理实验资源整合成一个“超级计算机”来共同完成同一个数据处理与分析任务成为可能。网格计算技术的出现和 WLCG 的建成和广泛使用直接促成了 Higgs 粒子的发现，在高能物理史上记下了重要的一笔。近年来，继网格计算之后，虚拟化技术和云计算技术的快速发展，正在为高能物理实验的科学计算输送更加强劲的计算能力。

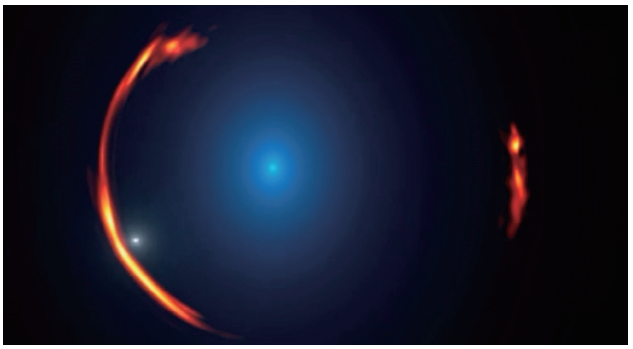
因此，纵观高能物理实验的发展历史，可以看出未来的高能物理实验仍需与先进计算技术紧密结合、互相促进，最终才能保证高能物理领域的长远发展。

科苑快讯

怎样发现暗星系

由于巨大质量造成的光线扭曲效应，天文学家辨认出一个主要由暗物质构成的矮星系，该星系属于一个距地球 40 亿光年的较大星团。该星系介于地球与另一个命名为 SPD.81 的更遥远星系（距离我们 120 亿光年）之间，遥远星系发出的无线电波在经过地球与其之间的星团时被扭曲成一个环，这就是“引力透镜”（gravitational lensing）效应。环的大小和形状有助于天文学家测量介于其间的星系的质量，大约相当于 1 万亿个太阳。

研究者说，由于环的大小和形状不能与单个星系波的弯曲程度进行精确匹配，未发现的恒星应该位于一个非常黯淡的以暗物质为主导的伴星系中，该星系围绕一颗较大质量的恒星运转。根据研究组在《天文物理学期刊》（*The Astrophysical Journal*）网络版上发表的计算结果，该卫星星系的质量约为太阳的 10 亿倍。研究组的分析结果使他们确定另一个看不见的矮星系（图中靠



近圆环的光点）的可能位置。一远一近两个恒星系的偶然对齐，帮助研究者进一步了解了钩成矮伴星系的暗物质类型，使天文学家能够更好地洞悉组成的宇宙无论是可见还是不可见的宇宙到底是由什么构成的。

（高凌云编译自 2016 年 4 月 14 日 www.sciencemag.org）