# 高能物理实验的数据密集型计算*

文 / 陈和生　陈 刚
中国科学院高能物理研究所 北京 100049

【摘 要】高能物理一直是信息技术发展的主要推动者之一。现代高能物理产生的海量数据对计算机技术提出巨大的挑战。为了应对这些挑战，国内外高能物理领域的科学家根据数据处理的特点建立新的计算平台用于传输、储存及分析处理PB量级的数据。文章介绍了现代高能物理实验及数据处理的发展，并描述了高能物理的计算模型以及以网格技术为代表的数据密集型计算平台；详细介绍了数据密集型网格平台在LHC实验、BESIII实验中的应用以及中国的数据密集型网格平台。并对云计算等新技术在高能物理领域的应用进行了展望。

【关键词】高能物理，大数据，数据密集型计算，网格，云计算

DOI 10.3969/j.issn.1000-3045.2013.04.010

## 1 高能物理简介

高能物理，又称为粒子物理，是物理学一个前沿分支。其科学目标是研究组成物质的最小单元及其相互作用规律。物质由原子组成，原子由原子核和电子组成。原子核由质子和中子组成。质子和中子则是由夸克组成的。粒子间的相互作用由中间波色子传递。现代高能物理研究的主要目标包括深入检验标准模型、探索超越标准模型的新粒子和新现象（更高的能量标度）。粒子物理与宇宙学和天体物理的交叉产生了交叉前沿学科——粒子天体物理。最新的天文观察结果表明，宇宙中存在暗物质和暗能量，分别占宇宙物质总量的23%和73%，而迄今为止标准模型描述的物质只占4%。高能物理面临巨大的挑战，正处于重大历史性突破的前夜。高能物理实验要求大型的科学实验装置，包括大型加速器和探测器。目前世界上最大的高能物理实验装置是在日内瓦欧洲核子中心（CERN）的大型强子对撞机（LHC）[1]，其主要物理目标是寻找希格斯（Higgs）粒子，超对称（Supersymmetric）粒子以及其他新物理现象。人们还在探讨建立更高能量的物理直线对撞机（ILC）或Higgs工厂，对LHC发现的新物理和新粒子进行更精确的研究。另外还有许多非加速

* 收稿日期：2013年4月15日

器物理实验正在探索超越标准模型的物理现象，包括粒子天体物理实验，宇宙线观测、中微子物理实验（如测量中微子质量的顺序，CP破坏……），寻找暗物质等等。这些实验寻找稀有事例，需要建设庞大的探测器，往往也是大科学装置。

高能物理是实验科学，实验验证理论并推动理论发展。理论物理学家利用实验观测的结果来验证理论，并提出推论或新的理论。新的理论又需要新的实验来验证。因此，实验是高能物理研究的基础，而实验的数据处理是物理分析研究的关键。高能物理实验的传统是根据其数据处理及物理分析的需求特点，结合信息技术建立自己的信息平台。该平台为数据的采集、存储、处理和分析，物理模拟及合作交流提供支撑。

## 2 高能物理实验和数据发展趋势

过去几十年高能物理实验的规模和复杂度都发生了巨大的变化。实验规模和复杂度的提高意味着数据量的增加和数据分析难度的增加。但是高能物理的计算模式基本不变。现代实验的数据采集系统对实验数据进行采集、甄别和快速过滤，形成实验原始数据。海量数据存储系统记录原始数据，用于后续数据分析处理。高能物理实验的精度依赖于数据的统计量，寻找稀有事例实验的数据量越来越大。原始数据在所谓的离线（相对于在线数据采集）计算系统中进行处理和分析。离线处理包括以下工作：（1）根据实验装置的特性和工作状态对原始数据的校准刻度；（2）根据粒子与实验探测器介质相互作用的性质对刻度后的数据进行事例重建，鉴别出具有明确物理意义的粒子及对应的物理参数；（3）对鉴别出来的粒子进行分类筛选，找出特定的物理事例并进行物理研究分析。所有这些计算都涉及大规模的数据处理，同时可能产生更多的重建数据和蒙特卡洛模拟数据。另外，随着实验中粒子能量的不断提高，数据的复杂度也越来越高。图1[2]和图2显示的是早期高能物理实验和最新实验中的粒子径迹数量的对比。

高能物理实验的发展使实验的规模和复杂度不断提高，实验数据产生、分析和处理对计算环境不断提出巨大的挑战。以20世纪末世界上最大的对撞机LEP（Large Electron Positron collider）为例，它的4个实验在1989—2000年整个实验期间积累的数据总共不到20TB。而最新的LHC对撞机实验每年采集的数据就达15PB以上。因此高能物理实验需要通过大规模的国际合作来进行数据分析。早期的高能物理实验的数据分析可以在一个数据中心内完成。20世纪80年代末之前，由于网络性能的限制，数据处理和分析也只能局限在一个数据中心中。从20世纪90年代开始，高能物理

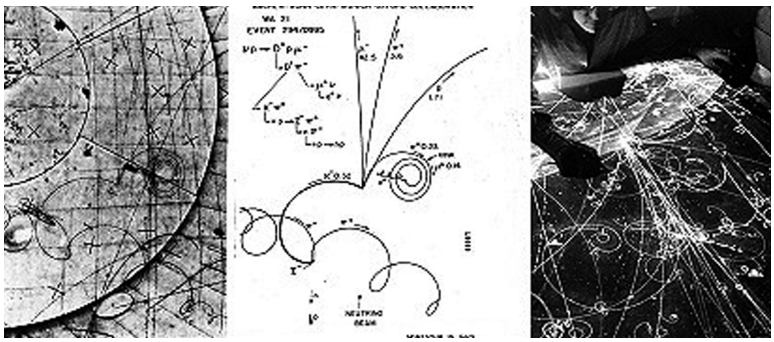

图1 20世纪70年代的实验，粒子在探测器（泡室）中的径迹

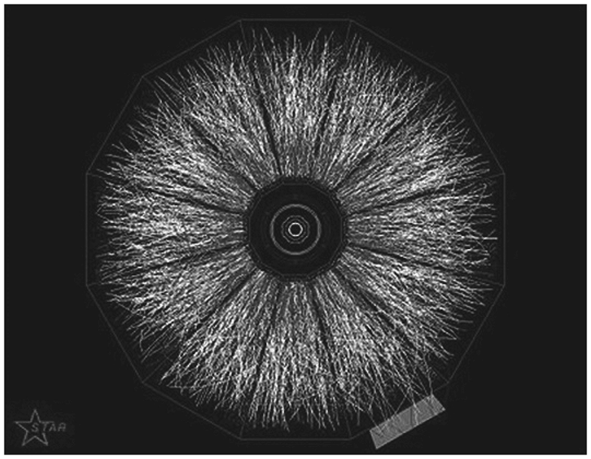

图2 布鲁克海文实验室STAR实验，金核与金核对撞产生的粒子在探测器的径迹

实验的规模出现了巨大的提升，来自世界各国的科学家联合参加同一个实验，实验的数据需要在多个数据中心进行分析处理。互联网的出现和普及为数据及计算资源的远程访问和共享提供了条件。WWW网页应运而生，不仅成为高能物理学家的基本交流手段，更带来一次深刻信息技术的革命，产生极为巨大的影响。随着包括对撞机实验、宇宙线实验等在内的新的高能物理实验的发展，数据的规模将进一步扩大。数据的处理不仅需要更高性能的网络的支撑，同时还需要最新的存储技术、计算机技术来应对新的挑战。

## 3 高能物理数据处理基础环境和相关技术

高能物理实验产生的数据经过高效处理和分析才能获得物理结果。因此，建立高性能数据处理基础环境是高能物理实验的重要工作之一，包括硬件环境和软件环境两大类。硬件环境包括数据存储、计算环境、网络环境三大资源。实验产生的海量数据需要安全可靠地保存起来，同时又能高效地访问。因此存储系统需要根据实验数据的规模及处理模式进行仔细设计，以满足实验需求。对撞机实验的对撞事例彼此没有关联，宇宙线粒子事例间也没有关联。因此高能物理数据的特点是海量，同时互不关联。尽管数据的格式有可能不同，但高能物理的实验数据都是以数据文件的方式存储的。目前每个数据文件的大小都在几个GB的量级。由于数据文件是没有关联的，因此可以启动一批独立的计算作业同时对数据文件进行分别处理。一般情况下，一个数据中心会同时提交上千个甚至上万个作业，这些作业会同时访问成千上万个数据文件，这就要求数据存储系统具有很高的聚合I/O吞吐能力以适应高并发访问请求。高能物理数据中心一般配备分布式的存储系统，如GPFS、ZFS、Lustre等等。现代高能物理实验的规模巨大，因此实验的数据常被分散到若干个数据中心进行存储和分析处理。考虑到数据安全，数据有时还采用异地备份。

就计算环境而言，高能物理数据绝大部分是以可分割及相对独立数据文件方式保存和处理的，因此并不需要大规模的内存共享的并行计算任务。高能物理的计算环境主要采用松耦合的计算集群系统。这种计算集群造价比较便宜，宜于升级。为了有效地利用遍布世界各地的实验合作单位的数据中心建立分布式的数据处理环境，高速网络是现代高能物理实验数据处理不可缺少的条件。数据中心之间至少需要Gbps级、10Gbps级甚至更高的网络带宽进行数据的传输和交换。

高能物理数据处理基础环境的软件部分主要包括资源管理系统和通用软件包两部分。资源管理系统用于对存储资源、计算资源及网络资源进行管理和调度。高能物理领域常根据数据处理的特点建立自己的数据格式以提高数据的存储及访问的效率和便利性。还开发针对大规模数据传输及广域网数据管理的系统，实现海量数据在数据中心之间的传输和管理。

尽管世界各地的高能物理实验的研究目标不同，实验也不同，但所涉及的物理过程具有很高的相似性。这为建立通用软件进行共享提供了可能，例如，物理学家开发的用于描述粒子相互作用的软件包GEANT4[3]。该软件包用来模拟粒子穿过介质时与物质发生作用的过程，从而帮助物理学家理解或预测实验产生的数据，为实验设计、数据分析处理提供依据。另外，物理学家还开发了各种通用的数字计算和物理分析软件包，如物理分析框架ROOT[4]。ROOT是一个面向对象的数据分析框架工

具，可用于大规模数据的分析处理和可视化。由于这些软件工具全部是开源的，几乎全世界的高能物理实验均采用这些软件包作为数据处理的基础，并在此基础上建立自己的数据处理系统。

高能物理数据的处理分析建立在上述的基础环境之上。数据处理涉及到以下任务及技术：

### 3.1 物理模拟

高能物理对撞产生的终态粒子（或者宇宙线）在探测器介质中的运动过程会与介质发生相互作用，从而留下时间、位置及能量沉积等信息。这些信息将被用来决定终态粒子的物理参数，如能量、动量、运动方向和粒子种类等等。由于粒子与介质的作用过程十分复杂，且具有随机性，因此必须用蒙特卡洛方法来模拟这些反应的详细过程，并数字化。

在高能物理实验装置设计阶段，需要对探测器做大量的模拟研究，以了解实验装置对终态粒子的响应，判断该装置能否满足物理目标的要求，并优化装置的设计。在探测器开始运行前，物理模拟数据还被用来检验数据分析软件的正确性和可靠性。

为达到足够的模拟精度，物理模拟必须产生与实际实验采集数量相当的事例，因此模拟过程也将产生海量的数据，而且是一个巨大的任务。

### 3.2 数据重建及物理分析

高能物理实验装置用来记录终态粒子穿过装置介质时留下的信息。每个信息记录点称为一个击中点或着火点（hit）。这些击中点的信息通过快速筛选和组合作为实验的原始数据记录并保存到存储系统中。原始数据需要通过筛选、模式识别及粒子鉴别才能变成具有物理意义的数据。这一过程叫做事例重建，产生的数据叫做重建数据。事例重建前先对探测器采集到的原始数据进行刻度和校准，然后进行径迹（及终态粒子在探测器中留下的轨迹）的寻找和拟合以及粒子的鉴别等。随着高能物理实验能量的不断提高，实验装置规模越来越大，采集的数据也越来越复杂，每个事例的数据信息个数甚至以百万计。重建过程因此非常复杂。

事例重建是高能物理数据处理最重要的环节，同时也是计算量最大的任务。事例重建同样可以在计算集群上进行。由于事例重建可能需要同时在数以千计的CPU上进行，每个事例重建的计算任务都需要从存储系统上快速地读取数据，计算结果产生的重建数据也需要输出到存储系统，因此事例重建需要能承受高并发访问高吞吐率的存储系统。高能物理除了采用前面提到的分布式存储系统以外，还根据特定的数据访问模式，设计开发了dCache、DPM等存储系统。这些系统都为高能物理数据处理提供了高性能的数据存储服务。

重建数据被用来进行物理分析，并获得最终的物理结果。物理学家通过交互式或者批作业的方式对数据进行分析，选取自己感兴趣的事例，从中寻找物理规律或新的发现。物理分析过程中需要读取重建数据，并对数据进行判选。这个过程同样需要存储系统的支撑。有时还需要可视化工具对事例进行展示，方便物理分析的进行。前面介绍的ROOT工具提供了优良的数据可视化手段。目前物理学家还在基于ROOT等工具开发3D展示的系统，为物理分析提供更好的可视化服务。

### 3.3 网格及分布式数据共享和处理

截至2012年底，仅LHC实验就积累了超过150PB的数据。未来几年世界高能物理的实验数据将超过1 000PB。这样的数据量需要超大规模的计算资源。网格技术把分布于全世界的存储、计算资源整合到一起，形成一个超高性能的通用的计算用基础设施。它提供的服务将包括：足够的计算和存储能力，用于数据的处理、模拟和分析；高速网络，用于各合作机构之间海量数据的传输；高效的资源互相访问工具，从而实现将大量的工作有效地分配给世界各地的合作成员。

国际高能物理领域建立了一系列分布式网格

计算系统，并联合形成面向高能物理等大科学的网格平台，其中包括欧洲的国际高能物理网络 WLCG[5]、美国的 TeraGrid 等等。网格平台为 LHC 等大型高能物理实验的数据处理及分析需求提供了保障。

## 4 典型案例

### 4.1 北京谱仪

北京正负电子对撞机 BEPC 和北京谱仪是国际上粲物理能区性能最好的高能物理实验装置。第三代探测器 BESIII[6]（图 3）的物理目标包括轻强子谱测量、粲偶素研究、粲介子物理、t 物理以及新物理探索。BESIII 从 2009 年开始采集数据，将至少继续运行 10 年。未来几年 BESIII 的数据规模将达到 10PB 以上。BESIII 实验的数据分析在实验停止运行后还将继续进行若干年，实验数据的生命期至少达 15 年以上。

BESIII 数据处理软件 BOSS（BESIII Offline Software System）是实验组根据探测器的特性以及物理目标自行开发的。该软件系统采用 C++和面向对象技术在科学 Linux（Scientific Linux）平台上进行开发。BESIII 数据处理及物理分析软件包括软件框架、模拟软件、刻度、事例重建和物理分析工具 5 个部分。

数据存储是 BESIII 实验的重大挑战之一。最经济高效的数据存储解决方案是支撑 BESIII 数据处理和物理分析的保障。BESIII 数据存储包括分级存储（HSM, Hierarchical Storage Management）系统和并行文件两部分。BESIII 分级存储系统称为 GRASS（Grid-enabled Advanced Storage System），是在欧洲粒子物理中心的 CASTOR 系统基础上开发的，包括 IBM TotalStorage 3854 磁带库和 LTO-4 磁带驱动器组成的磁带库系统，以及磁盘阵列组成的磁盘池以及 GRASS 存储管理系统 3 部分。并行文件系统基于 Lustre 文件系统进行优化改进，并采用低端硬件平台建立。该系统对 Lustre 的稳定性和并发访问性能等进行了改进。到 2012 年底，并行文件系统的容量达 3PB，并发访问性能达到 25GB/s 以上。

BESIII 实验数据处理的另一个重大挑战是数据共享和分布式处理。BESIII 实验是大型国际合作，需要在国际合作成员之间进行高效及时的数据共享，同时海量数据集中在一个数据中心进行处理本身就不是理想的方案。因此 BESIII 采用了网格技术。

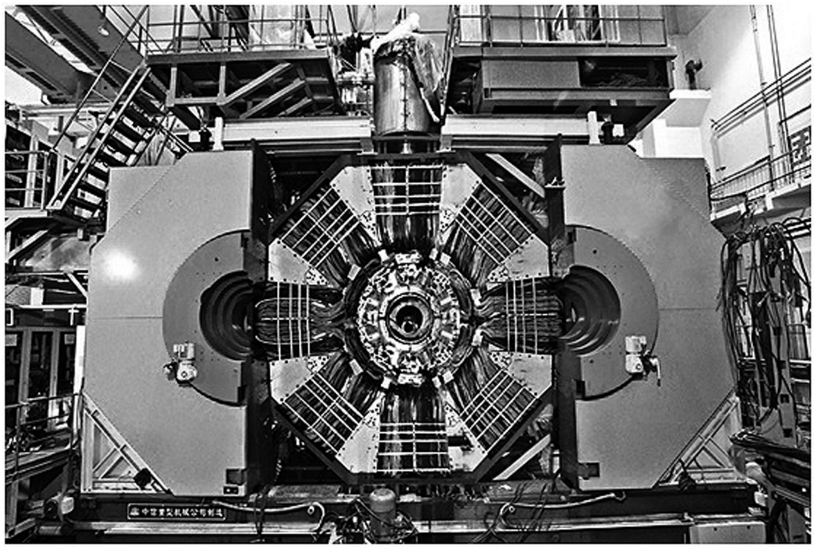

图 3 北京谱仪 BESIII

BESIII 网格平台由高能物理所的一个中心站点和国内外的若干格卫星站点组成。网格平台采用 gLite（未来将升级成 EMI）为中间件，同时也可以与中国国家网格的

GOS中间件实现互操作。计算任务可在网格站点之间全局调度。网格平台的数据传输管理采用DIRAC系统，可高效智能地实现站点间的数据传输。BESIII网格系统将计算任务和数据调度到俄罗斯、美国、德国及国内的数据中心，每年完成的计算作业达数百万个，为BESIII数据处理提供重要支撑。

### 4.2 LHC实验

LHC位于日内瓦的CERN。它建造在周长为26.66公里的地下隧道里。两束能量各为7TeV的质子在LHC中进行对撞。来自全世界超过6 000名科学家参加LHC的4个主要实验：ALICE、ATLAS、CMS、LHCb（图4）。这4个实验将探索粒子物理学最前沿的课题，包括寻找质量起源的Higgs粒子以及超对称粒子等。基本粒子和Higgs粒子的相互作用使基本粒子具有质量。2011年12月LHC的CMS实验和Atlas实验宣布观察到类似于Higgs的粒子。随着数据的积累，Higgs粒子的实验证据得到了进一步确认。LHC的第二个重要目标是寻找暗物质。科学家希望通过LHC能发现暗物质粒子。LHC将在人类对物质结构的认识方面实现一次重大跨越。

中科院高能物理所、原子能研究院、北京大学、清华大学、南京大学、山东大学、中国科技大学、华中师范大学等分别参加了LHC的4个实验。

LHC对撞机和4个实验于2009年投入运行，每年将产生约15PB的原始数据。实验将运行20年以上，储存这些数据并进行分析处理，这对计算系统是一个巨大的挑战。实验物理分析需要至少20万个CPU和海量的数据存储系统。由于数千个物理学家分布在世界各地，为了方便高效地进行物理数据分析研究，LHC采用分级式（Tier）的计算平台，将实验数据复制到各地区数据分析中心。这种解决方案就是WLCG（图5）。LHC实验决定采用相对便宜的硬件来建立其计算环境，而不采用昂贵的高端数据服务器和计算机。这种方式和Google采取的策略相似。WLCG所谓的分级结构由0—2级等规模不同的计算中心组成。各地区的一级中心(Tier-1)与CERN的零级中心（Tier-0）之间至少需要10Gbps的网络带宽。二级中心(Tier-2)与一级中心之间的网络则至少需要2.5Gbps。零级中心负责数据的备份及向其他中心的数据分发，一级中心往往由参加LHC实验的成员国建立，二级中心则由规模较大的研究机构建立。LHC实验能够利用该网格系统存储和分析数据。WLCG在全球的网格站点达200余个，大规模网格系统的一个重要挑战就是数据安全问

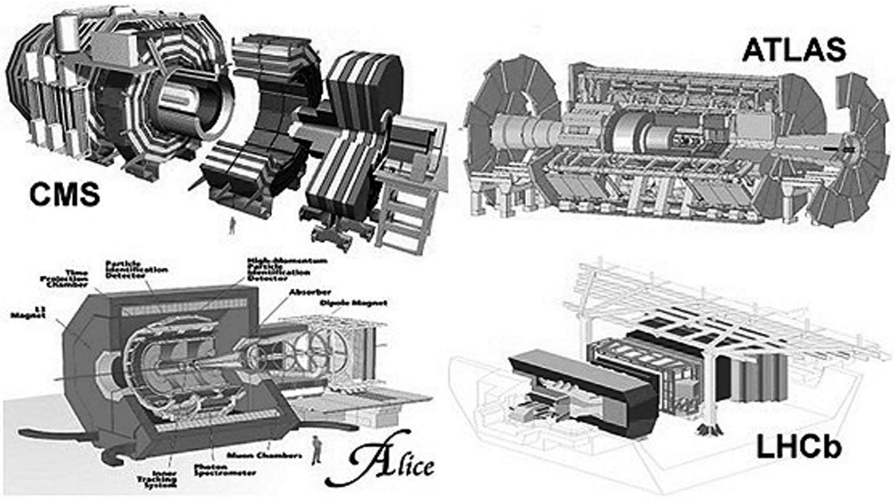


图4 LHC的4个主要实验：Alice、ATLAS、CMS和LHCb

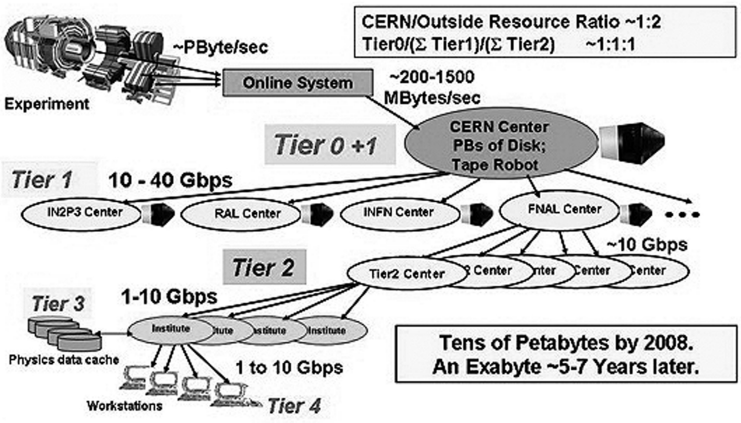


图5 WLCG分级结构

题。WLCG不能依赖于防火墙系统，因为这将成为大规模数据传输的瓶颈，因此采用数字身份认证和授权的手段来保证数据不被非法访问。

WLCG作为世界上最大的网格平台之一，目前装备了超过25万个CPU核及150PB的存储资源，每年完成数亿CPU小时的计算任务，为LHC实验的数据分析处理提供了不可或缺的支撑，特别是为Higgs粒子的发现做出了巨大贡献。

## 4.3 WLCG中国站点

2006年高能物理所代表ATLAS和CMS中国合作组与CERN签署协议，加入WLCG的建设和运行，支持ATLAS和CMS实验的海量数据处理。在中科院知识创新重大项目的支持下，于2008年在高能物理所建立了WLCG网格平台二级站点。该网格站点由约1 600个CPU核组成计算资源，640TB的磁盘组成存储系统。计算资源采用刀片式服务器。磁盘存储采用廉价的硬件设备，配备dCache和DPM作为存储管理系统。该系统的优点是性能好，可靠性高且易管理。网格平台通过中国科技网建立了到欧洲和北美的高速网络带宽。与欧洲的网络连接采用ORIENTplus链路，与美国的网络连接采用Gloriad链路。每年与欧

**Tier-2 Availability and Reliability Report**
**for VO ops (WLCG_CREAM_LCGCE_CRITICAL)**

Federation Summary - Sorted by Reliability — January 2012

Data from Nagios and ACE
https://twiki.cern.ch/twiki/pub/LCG/GridView/Ace_Service_Availability_Computation.pdf

Availability = Uptime / (Total time - Time_status_was_UNKNOWN)
Reliability = Uptime / (Total time - Scheduled Downtime - Time_status_was_UNKNOWN)
HS06 : Installed capacity of the site measured in HEPSPEC06 (HS06)
Reliability and Availability for Federation - Weighted average of all sites in the Federation based on installed capacity(HS06)

Colour coding : N/A | < 30% | < 60% | < 90% | >= 90%

| Federation | Reli-ability | Avail-ability | Federation | Reli-ability | Avail-ability |
|---|---|---|---|---|---|
| PT-LIP-LCG-Tier2 | 100 % | 100 % | T2_US_Purdue | 99 % | 99 % |
| CN-IHEP | 100 % | 100 % | BE-TIER2 | 98 % | 98 % |
| KR-KISTI-T2 | 100 % | 100 % | FR-IN2P3-CC-T2 | 98 % | 98 % |
| NO-NORGRID-T2 | 100 % | 100 % | CA-EAST-T2 | 98 % | 94 % |
| DE-DESY-LHCB | 100 % | 100 % | FR-IN2P3-IPHC | 98 % | 93 % |
| FR-GRIF | 100 % | 100 % | AU-ATLAS | 98 % | 98 % |
| SI-SiGNET | 100 % | 100 % | FR-IN2P3-LPSC | 97 % | 97 % |
| US-NET2 | 100 % | 100 % | US-MWT2 | 97 % | 97 % |
| FR-IN2P3-CPPM | 100 % | 100 % | RO-LCG | 97 % | 97 % |
| FR-IN2P3-LAPP | 100 % | 92 % | HU-HGCC-T2 | 97 % | 96 % |
| US-LBNL-ALICE | 100 % | 100 % | FR-IN2P3-SUBATECH | 97 % | 96 % |
| US-LLNL-ALICE | 100 % | 100 % | UK-ScotGrid | 96 % | 96 % |
| CZ-Prague-T2 | 100 % | 99 % | IT-INFN-T2 | 95 % | 94 % |
| T2_US_Florida | 100 % | 100 % | FR-IN2P3-LPC | 95 % | 95 % |
| JP-Tokyo-ATLAS-T2 | 100 % | 100 % | US-WT2 | 95 % | 94 % |
| T2_US_Caltech | 100 % | 100 % | SE-SNIC-T2 | 94 % | 94 % |
| ES-CMS-T2 | 100 % | 99 % | UK-SouthGrid | 93 % | 93 % |
| GR-Ioannina-HEP | 100 % | 95 % | DE-DESY-ATLAS-T2 | 93 % | 93 % |
| T2_US_Wisconsin | 100 % | 100 % | ES-ATLAS-T2 | 93 % | 93 % |
| BR-SP-SPRACE | 100 % | 62 % | ES-LHCb-T2 | 93 % | 92 % |
| UK-London-Tier2 | 100 % | 99 % | TR-Tier2-federation | 92 % | 92 % |
| CA-WEST-T2 | 100 % | 99 % | UK-NorthGrid | 92 % | 92 % |
| T2_US_MIT | 100 % | 100 % | DE-FREIBURGWUPPERTAL | 91 % | 91 % |
| US-SWT2 | 100 % | 100 % | RU-RDIG | 91 % | 90 % |
| T2_US_UCSD | 99 % | 98 % | DE-MCAT | 91 % | 90 % |
| TW-FTT-T2 | 99 % | 98 % | IN-INDIACMS-TIFR | 90 % | 84 % |
| UA-Tier2-Federation | 99 % | 99 % | DE-DESY-RWTH-CMS-T2 | 90 % | 88 % |
| FI-HIP-T2 | 99 % | 99 % | IL-HEPTier-2 | 90 % | 89 % |
| T2_US_Nebraska | 99 % | 99 % | EE-NICPB | 88 % | 88 % |
| US-AGLT2 | 99 % | 97 % | KR-KNU-T2 | 86 % | 86 % |
| PL-TIER2-WLCG | 99 % | 99 % | AT-HEPHY-VIENNA-UIBK | 77 % | 77 % |
| PK-CMS-T2 | 99 % | 98 % | DE-DESY-GOE-ATLAS-T2 | 63 % | 43 % |
| CH-CHIPP-CSCS | 99 % | 99 % | IN-DAE-KOLKATA-TIER2 | 1 % | 1 % |

Page 3 of 10

图6 中国网格站点(CN-IHEP)的运行水平位列世界前列

洲及北美之间交换3PB以上的数据。2013年初在中国科技网的帮助下，对与欧洲的网络宽带进行了大规模升级，目前的国际数据传输性能达到4.6Gb/s以上。高能物理所还建立了CA安全认证授权中心。该授权中心是国内唯一通过欧洲网格安全授权组织（EUGridPMA）和亚太网格安全授权组织（APGridPMA）的双重认证的授权系统，为高能物理等领域使用网格系统的个人签发CA证书，同时还为网格平台的主机以及服务签发CA证书。多年来，中国网格站点在全球近200个网格站点中运行水平一直处于世界领先地位（图6），特别是被ATLAS国际合作组评为Leadership站点。该网格站点每年提供超过1 200多万CPU小时的计算服务，完成550余万个计算作业，处理的数据超过3PB，为ATLAS、CMS实验的物理分析（尤其是对2012年7月Higgs玻色子的重大发现）做出了重要的贡献。

网格平台不仅提供计算和数据处理服务，同时还帮助实现了人力资源的共享。2009年6月，高能物理所为CMS建立了CMSROC@Beijing区域运行中心。CMSROC@Beijing是继美国费米实验室和德国电子同步加速器研究所之后的第3个区域运营中心。这是CMS首次将远程运行从欧洲、北美扩展到了亚洲。3个运行中心分别位于3个不同的时区，每个运行中心值班8小时，这样就实现了24小时轮班制。这种轮班制可以有效地保证CMS实验的顺利进行。区域运行中心帮助中国物理学家更方便地参与CMS实验的研究活动。

## 5 总结和展望

随着计算机及网络技术的不断发展，高能物理数据处理的技术与手段也在不断发展。在LHC实验建造初期，单个数据处理中心的CPU能力、磁盘容量都不能满足LHC实验海量数据处理的要求，因此LHC建立了分布式的网格平台将数据处理的任务分发到全世界近200个数据中心。由于当时网络带宽的限制，LHC的数据处理任务的分发采用的是以数据为中心的模式，即将计算任务提交到存放有相应数据的数据中心进行运行。最近几年，网络性能大幅提升，数据中心可以用10Gbps甚至数十Gbps的高速网络进行连接。因此WLCG将数据处理任务分发改成了以CPU为中心的模式，即实时地将数据传送到CPU空闲的数据中心，并在该数据中心进行处理。这为计算任务的调度分发提供了更大的灵活性。

WLCG提供了数据密集型计算的一个成功范例。WLCG实际上已经为许多其他领域的数据密集型计算提供了强有力的平台，为生物医学、天体物理、地质地理、气象研究等非高能物理领域的科学计算提供了广泛的支持。以中国的站点为例，该站点不仅为LHC实验提供服务，还为中科院大连化学物理所的蛋白质结构研究、中科院大学的地球动力学研究、国际病毒药物筛选Wisdom项目以及国际核磁共振及结构生物学WeNMR项目提供计算及数据处理服务，为这些科学研究做出了重要贡献。

计算机及网络技术的发展似乎在弱化分布式网格平台的必要性，但是高能物理实验的规模不断提升、实验数据量飞速增长，海量数据的处理需要新技术的支持。下一代高能物理实验，如未来直线加速器实验，大型宇宙线观测实验等将产生更大规模更复杂的数据。这些都将对计算技术提出新的挑战。分布式的计算平台为高能物理的国际合作提供了更大的方便，同时也大大降低了实验成本。因此分布式计算平台仍将是高能物理数据处理和计算的重要模式。

云计算是当前热门的计算模式。但是由于高能物理数据量太大，且基本上是一次写多次读，采用商业云平台需要的网络开销太大。根据LHC等实验的测试和评估表明，采用商业云的成本要高于目前的网格平台。但虚拟化技术为跨平台的计算任务调度和资源整合提供了技术条件。云计算技术在提高资源利用率、灵活的可伸缩性及可管理性方面表现出了巨大的优势，吸引了包括高能物理在内的多个领域开始测试和应用。CERN启动了虚拟机项目CernVM[7,8]，并在此基础上发起

LHC云计算项目[9]，为LHC提供虚拟化的应用环境。同时，CERN还启动lxcloud项目[10]支持批处理计算服务，以提高资源利用率并简化管理。高能物理所计算中心在对高能物理实际应用需求进行详细分析后，认为只要能够满足需求的技术都是好技术，因此并没有简单地抛弃已有技术，而是结合现有的技术优势，包括网格计算、志愿计算、海量存储、下一代互联网及网络安全等，在云存储系统、虚拟集群系统、BESIII云计算系统及云安全等方面展开研究和应用。计算中心在现有海量存储技术基础上，基于实际需求设计与开发了一套云存储系统HepyCloud，轻松管理PB级乃至数十PB的存储空间。计算中心结合志愿计算[11,12]、虚拟化技术以及网格计算等技术，启动BESIII弹性云计算项目，不仅将BESIII计算任务分布到合作单位的计算系统，还将任务分发到互联网上的个人计算机中运行。而对于BESIII的用户来说，仍使用原有的作业提交方式，而不用关心作业被分发到本地集群、WLCG网格站点或者中国国家网格CNGrid站点上，还是个人计算机上执行。

应当看到，云计算概念目前还没有统一认识和定义，每个行业都从自身的角度来看待，还在不断的发展和完善。新的技术还将不断涌现。但是无论如何，技术是为应用而服务的，应用始终是推动技术发展的源动力。高能物理的数据处理及计算平台仍会借助于新的信息技术，同时也将推动数据技术的发展。

## 参考文献


1 大型强子对撞机LHC：http://lhc-machine-outreach.web.cern.ch/lhc-machine-outreach/introduction.htm.

2 Bubble chamber: D meson production and decay，CERN-EX-68681.

3 GEANT4, a toolkit for the simulation of the passage of particles through matter: http://geant4.cern.ch/.

4 ROOT: http://root.cern.ch/drupal/content/about.

5 WLCG: http://lcg.web.cern.ch/LCG/

6 北京谱仪：http://bes3.ihep.ac.cn/

7 Predrag Buncic et al. CernVM, http://cernvm.cern.ch/cernvm.

8 Buncic P et al. CernVM—a virtual appliance for LHC applications, Proceedings of Science，PoS（ACAT08）012, 2009.

9 Segal B, Buncic P et al. LHC Cloud Computing with CernVM, Proceedings of the 13th International Workshop on Advanced Computing and Analysis Techniques in Physics Research. February 22-27, 2010, Jaipur, India.

10 Tony Cass, Sebastien Goasguen et al. The batch virtualization project at CERN. EGEE09 conference, Barcelona.

11 中国科学院志愿计算网站：http://casathome.ihep.ac.cn.

12 David P. Anderson and Gilles Fedak, The Computational and Storage Potential of Volunteer Computing. Sixth IEEE International Symposium on Cluster Computing and the Grid, 73-80.


## Data Intensive Computing in High Energy Physics


Chen Hesheng　Chen Gang

（Institute of High Energy Physics, Chinese Academy of Sciences，Beijing 100049，China）



**Abstract** High energy physics (HEP) has always been a pioneer to develop information technologies. Modern HEP creates gigantic data sets which lead the huge challenges to the computer sciences. Scientists of HEP community developed the state-of-art computing platform to distribute, store and process data in PB scale. This report describes the evolution of high energy physics experiments and its computing technologies. The comput-

ing models and grid computing as the examples of data intensive computing platform are discussed in details. This report also introduces the application of grid computing in the LHC and BESIII experiments. The Chinese data intensive grid systems are are reported. The prospect of next genertion technologies such as cloud computing is discussed.

**Keywords** high energy physics, big data, data intensive computing, grid computing, cloud computing

**陈和生** 中科院院士，中科院高能物理所研究员，北京正负电子对撞机国家实验室主任。1998—2011年任中科院高能物理所所长，历任中国物理学会副理事长、中国高能物理学会理事长、国际未来加速器委员会委员、国际高能物理计算技术委员会委员、亚洲未来加速器委员会主席等职。长期从事粒子物理实验，对发现胶子喷注、检验电弱理论、精确测定电弱参数和中微子代数等重大研究成果做出了重要贡献。1995—1997年在北京主持阿尔法磁谱仪大型永磁体系统研制，该磁体于1998年搭乘航天飞机成果进行首次飞行，成为人类送入宇宙的第一大型磁体，并于2011年送至国际空间站长期运行。2004—2009年主持北京正负电子对撞机重大改造工程（BEPC Ⅱ）建设。现主持中国散裂中子源工程建设。E-mail: chenhs@ihep.ac.cn

**陈 刚** 男，中科院高能物理所研究员、计算中心主任。1991年开始参加由丁肇中教授领导的大型粒子物理L3实验和阿尔法磁谱仪（AMS）实验，负责数据处理、物理分析及实验装置的设计建造。在从事粒子物理实验研究工作的同时，还负责高性能计算环境的研究工作。2004年开始在国内建立了高能物理网格计算环境。2008年负责建立国内数据密集型网格平台，为多个大规模科学研究项目提供计算平台服务。目前的主要研究方向包括高性能计算、高性能存储及网格技术。E-mail: GangChen@ihep.ac.cn

（接481页）

than ten years. It has been boosting a range of scientific innovations in CAS. This paper analyzes the trend of the international cyberinfrastructure, reviews the development of the CAS cyberinfrastructure and its applications, and presents its development opportunities, challenges, and the future direction.

**Keywords** cyberinfrastructure, applications, science and innovation

**南 凯** 中科院计算机网络信息中心副主任，博士，研究员，博士生导师。1974年出生。主要研究方向为分布式系统、网络协同工作环境。E-mail: nankai@cnic.ac.cn