

数据密集型网格平台示范站点与应用

- 陈刚 中国科学院高能物理研究所计算中心 北京 100049 Gang.Chen@ihep.ac.cn
- 程耀东 中国科学院高能物理研究所计算中心 北京 100049

摘要：

欧洲大型强子对撞机LHC和北京正负电子对撞机BEPCII等一代高能物理实验已经于两年前开始运行，这类实验每年会产生总量超过15PB的数据，分析处理这些数据并进行物理前沿课题的研究对计算机是一个巨大的挑战。通过十多年的发展，网格计算已经成为高能物理数据分析与研究的重要工具。由世界各国科研机构共同建设的高能物理网格平台不仅成功地为国际高能物理提供分布式高性能计算服务，同时也成为许多非高能物理领域的科学计算平台。中国数据密集型网格平台由中国科学院、国家自然科学基金委及科技部共同支持建设，其资源包括由1600个CPU内核组成的计算集群和640TB磁盘组成的高性能存储系统。网格平台与欧洲及北美建立了高速网络连接并加入国际高能物理网格系统，并支持高能物理、天体物理与宇宙线物理、生物信息以及地球动力学等众多领域。

关键词：网格，数据密集型计算，高能物理，中间件，计算平台

1. 引言

新一代高能物理实验越来越复杂，实验数据规模越来越大。以欧洲粒子物理中心（CERN）大型强子对撞机LHC^[1]为例，每年产生的实验数据将达到15PB以上；北京正负电子对撞机（BEPCII）^[2]上的大型实验BESIII^[3]在未来几年内将产生6PB以上的数据。这些实验的数据处理和物理分析需要一个大规模的计算平台。网格系统作为分布式的计算环境为高能物理的数据处理与计算提供了优秀的解决方案。

国际高能物理网格WLCG^[4]将世界各地的计算中心和网格资源连在一起，用于处理和分析大型强子对撞机等高能物理实验产生的海量数据。该网格系统集成世界上几种主流的网格中间件并在其上部部署物理软件，从而为相关的高能物理实验提供高性能的计算服务。WLCG不仅成功地为高能物理提供服务，同时还为许多非物理的应用提供服务，其中包括生物信息、环境科学、地质地理等等。

中国科学家积极参加了大型强子对撞机的国际合作，加入国际高能物理网格平台并获得物理数据、计算资源的共享是产生物理结果的必要条件。中国的北京正负电子对撞机等一批大科学项目同样也需要网格计算平台的支撑。在中国科学院、国家自然科学基金委及科技部的支持下，高能物理研究所根据中国的特点建立了数据密集型网格平台并加入到国际高能物理网格合作，该网格系统已成为国际高能物理网格重要的二级站点，不仅为大型强子对撞机、北京正负电子对撞机等高能物理实验提供高性能的计算支持，同时也为天体物理、生物信息和地质地理等领域提供优质的服务。

2. 高能物理网格

2.1 WLCG国际高能物理网格

大型强子对撞机LHC是目前世界上最大的基础

本工作得到中国科学院知识创新工程重大项目（项目编号KJ CX1-YW-17），中国科学院信息化专项e-Science示范项目（项目编号INFO-115-D01-Z002），国家自然科学基金（项目编号90912004）和科技部863项目（项目编号2006AA01A120）的支持。

科学研究实验装置，2010年3月开始正式运行。前所未有的高能粒子对撞将帮助人类对自然进行全新的认识。LHC对撞机上共有4个大型高能物理实验，每年将产生15PB以上的数据。来自世界各国的数千名科学家参与这些实验的数据分析和物理研究。CERN发起建立了WLCG国际高能物理网格系统，目标是建立一个分布式的大规模数据存储和分析计算基础设施，用于支撑科学家的物理研究。

国际高能物理网格平台由欧盟的EGEE (Enabling Grids for E-sciencE) [6]网格系统、美国的OSG (Open Science Grid)[6]网格系统以及其他若干国家和地区的网格系统组成，并集成了不同的网格中间件和信息技术。目前，高能物理网格已经形成由40个国家和地区参与、250余个计算中心组成的大型网格平台。该网格平台采用分级式(Tier)的计算平台，根据任务的不同将网格平台分成零(Tier-0)到三级(Tier-3)的四个层次的网格中心或站点。零级中心最大，负责数据存储和初步分析。一级中心负责地区（如亚太地区）或国家中心的数据分发和大规模的数据处理。二级中心一般在一个国家内提供计算服务。三级中心为科研机构内的小型网格平台，不对外提供共享。

3. 高能物理网格关键技术

3.1 网格中间件

高能物理网格的关键服务包括安全服务、信息系统、作业调管理、数据管理以及运行等等。网格系统的核心组件是网格中间件，用于管理广域网上网格资源的发现、访问、分配、监控和统计等服务。过去十余年来，世界各国开发了多种不同的网格中间件，包括 Globus[7]，gLite[8]，GOS[9]，OSG，NAREGI[10]等等。以gLite为例，该中间件由EGEE开发，并广泛地应用于高能物理等领域，其总体结构如图1所示，该中间件系统包括若干组建，其中作业调度管理系统WMS (Workload Management System)接受用户的作业请求，从信息系统BDII中查询网格资源的状态，并为作业分配网格资源；资源信息系统BDII (Berkeley Directory Information Index)作为网格资源信息查询接口，提供信息查询服务；数据文件目录系统LFC (LCG File Catalog)提供数据文件的物理文件名与逻辑文件名的映射；计算单元CE (Computing Element) 是网格站点的本地资源，包括了计算集群、本地作业管理系统以及面向网格的作业调度接口；存储单元SE (Storage Element)为网格系统提供本地存储资源服务，它支持网格环境的数据传输和数据管理。

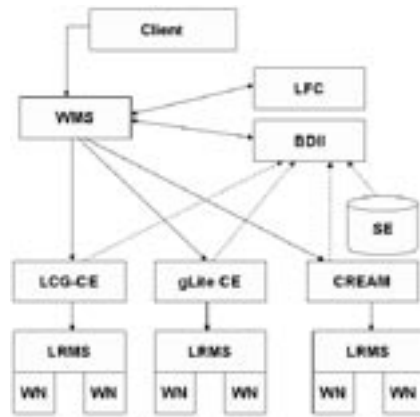


图1 gLite中间件结构

3.2 安全服务系统

对于分布式计算环境来说，安全是非常重要的因素。安全服务包括认证、授权和审核等，实现对用户、系统和服务的鉴别并决定对资源和服务的访问权限。安全服务还为安全事件的剖析提供信息。

3.2.1 网络安全基本架构

网格信息的传输在加密的情况下进行的，网络安全基于密钥的加密和解密。每个实体（个人、主机和服务）都持有一个妥善保管的私钥，并与一个公钥相对应。利用私钥加密的数据可以用公钥解密，而用公钥加密的数据也可以用私钥解密。私钥还用于对数据进行加密签名，接收数据方用公钥对签名进行解密并比对，以确认数据来自真实的发送方且未被篡改。

实体的身份认证利用电子证书来实现。网格技术中使用的证书是X.509格式的。证书的作用是保证每个实体的标识名是唯一的。证书由证书签发机构（Certificate Authority, CA）签发。网格系统的证书签发机构由不同地区的学术机构负责运行，为本地区的用户提供证书服务。证书的签发采用一定的手段对证书申请的身份进行确认从而决定是否签发证书。

3.2.2 高能物理研究所证书签发机构（IHEP CA）

IHEPCA于2004年运行，该证书签发机构通过了国际网格信任联盟（International Grid Trust Federation, IGTF）[11]认证，是国内最早的网络安全证书签发机构，并为国内物理、生物、地球科学等领域的用户签发证书。

IHEP CA采用现在电子商务通用的认证方式-PKI体系。在网格环境中，实体身份由X.509证书唯一标识，而证书的产生和发布则由CA来完成。为了与国际上其他证书签发机构互相认证，IHEP CA使用了包含所有签发机构的根证书（CA's root certificates）、撤销证书列表(Certificate Revoked List

, CRL)以及签发规程等信息的列表。IHEP CA系统结构如图2, 其中CA服务器用于签发和撤销证书, 是整个系统的核心。注册中心(Registration Authority, RA)处理所有来自用户的请求, 包括接收和批准请求、生成私钥、删除错误的请求和给用户回复邮件等等。安全服务器是整个系统的前端服务器。网格用户可以从安全服务器请求、查询和下载证书及撤销证书列表。用户和安全服务器之间的通讯遵循安全套接层SSL协议。数据库服务器保存证书申请请求、证书撤销请求、根证书等。

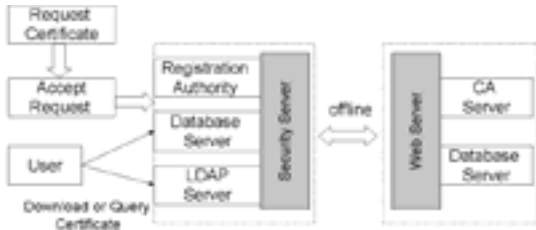


图2 IHEP CA体系结构

3.3 作业管理

作业管理系统(Workload Management System, WMS)由一组网格中间件组成, 用于网格资源环境中调度和管理作业。作业管理系统根据作业需求选择最适合的网格计算资源, 在此过程中, 需要考虑存储、网络带宽等因素, 对于不同的作业请求还需要考虑采用不同的调度策略。

3.3.1 典型作业管理系统的结构

作业管理的典型过程分为三步。第一步, 作业管理系统从用户的客户端接收作业提交请求并保存在作业队列中; 第二步, 作业管理系统启动一个匹配进程并根据作业请求寻找最佳的资源; 最后, 作业管系统将作业提交给选定的计算资源。

3.3.2 作业调度策略

高能物理网格采用基于引导代理(pilot agent)的作业管理系统, 其典型结构如图3所示。引导代理先于计算作业用传统的网格作业管理系统被提交并运行在计算资源上, 当引导代理在计算节点上运行时, 它首先检查该计算资源的状况, 如果发现合适的计算资源, 它立刻和基于引导代理的作业管理系统联系, 并将用户提交的作业拉到本地计算资源上来, 引导代理必要时还负责数据的传输和失败作业的恢复等等。

拉模式的作业调度最大的特点是性能好, 在高能物理网格中已经得到广泛的应用。例如, 高能物理网格的ATLAS应用每个月在世界上的100余个网格站点上执行约两千万个作业。

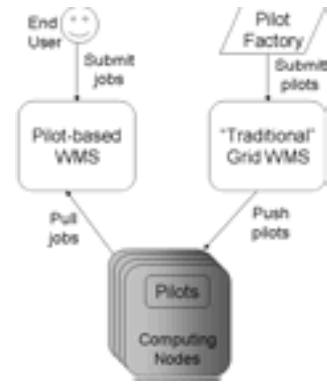


图3 基于引导代理的作业调度

3.3.3 跨网格平台的作业管理

高能物理等科学项目的规模巨大, 使用的网格系统需要在国际范围内进行部署, 因此需要跨网格平台的作业管理系统, 网格互操作是必须解决的问题。网格互操作的方式有多种, 包括应用驱动方式、网关方式和网格接口标准方式等等。

中国国家网格CNGrid^[12]和欧洲EGEE是两个大型网格平台, 高能物理等应用需要部署在这两个不同的网格平台上。但是CNGrid采用的中间件GOS和EGEE采用的网格中间件gLite不同, 因此高能物理研究所设计开发了一个作业管理系统, 将两个网格平台整合到一起。该作业管理系统如图4所示, 用户通过GUI或命令行方式发起的作业请求由作业管理系统接收以后根据要求被分成若干小作业并发送给作业提交模块, 作业提交模块调用相应的插件将作业提交给不同的网格平台。

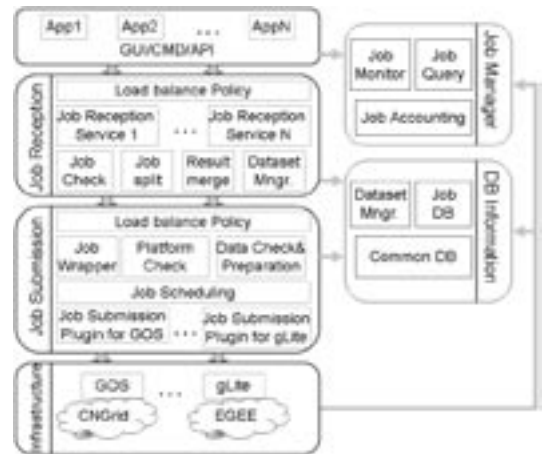


图4 跨网格平台作业管理系统

3.4 网格数据管理

3.4.1 数据管理概述

分布式数据管理是网格技术面临的最重要的一个挑战。在高性能物理网格上, 需要移动的数据量巨大, 数据管理必须具备高可控性, 以保障数据的正

确处理和安全存放。此外，数据管理还需要具备健壮的数据存储和副本机制，高容错性，智能性，从而减少人工干预。这些特性是对数据管理技术的巨大挑战。

在大规模的高能物理网格的生产系统中，数据传输的可靠性和健壮性非常重要。以gLite为例，中间件中的文件传输服务（FTS）为各个站点之间的数据流动提供了可靠、易管理的传输通道。FTS本身不是传输协议，它使用了已有的传输协议，如GridFTP, http, ftp等。FTS提出并且实现了传输通道的概念，使得FTS能掌控网格项目中的全局数据传输。目前，高能物理网格的零级与一级站点都部署了FTS的服务，从而管理站点之间每天几百个TB的文件传输。

3.4.2 BESIII数据管理系统

BESIII是北京正负电子对撞机上的大型粒子物理实验，实验在数据管理方面具有自己的特殊需求，因此需要开发自己专门的数据管理系统。该数据管理系统基于数据集（一系列的文件集合），对数据集的定义和存储位置被存储在数据集记录系统中。所有的数据和目录都位于以高能物理研究所的中心站点，一系列分布式的代理被运行在隶属于中心站点的所有下级站点，与数据管理服务进行交互，如图5所示。

中心站点的存储系统采取了混合模式：基于磁带库的海量存储系统GRASS和高数据访问性能的并行分布式文件系统CPFS的结合。GRASS作为一个后端系统，主要管理磁带库和数据在磁带库和磁盘服务器之间的迁移。CPFS提供了高性能的数据访问能力。到目前为止，高能所中心站点已经具有大约1600TB的CPFS磁盘存储空间，数据访问的聚合带宽能达到20GB/s以上。这种混合模式具有两大优势：一是减少了对后端系统GRASS的数据访问压力，降低了管理难度；二是前端的CPFS系统提供了高速的数据访问吞吐率，满足了高速数据访问的需求。此外，CPFS系统能被挂装到本地，类似于本地文件系统，为用户进行数据访问操作提供了一个易用的接口。

数据管理系统中的数据流动都是由订阅系统自动触发的。下属站点的管理员通过数据管理系统接口提交数据订阅请求。运行在下属站点的代理会检查数据管理系统中是否存在需要执行的任务，如果存在，则会从下级站点触发数据的传输。当数据传输完成后，该代理将数据集以及文件的状态报告给数据集记录系统。因此，BESIII网格中的所有数据文件都被登记在目录服务中，以便网格作业对文件的正确定位与访问。

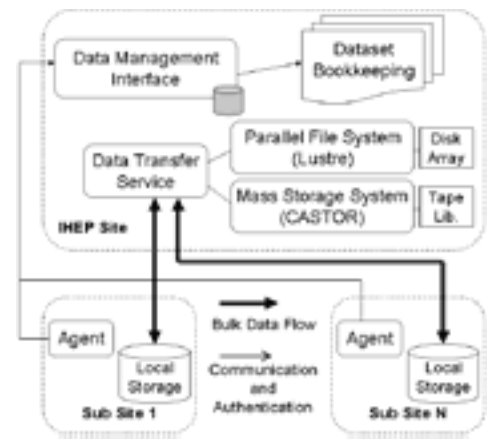


图5 BESIII数据管理

3.5 运行与维护

高能物理网格是一个高度异构的全局分布式系统，面临了很多技术的挑战。一个好的中间件是网格项目成功的关键。此外，高效的操作运行并保障所有网格服务的稳定也至关重要，一旦发生问题，需要及时被发现并解决。网格监控系统正是用来检测网格服务稳定状态的系统。

除了监控系统，规则与规范也同样重要。在加入高能物理网格之前，每个站点都需要与高能物理网格合作项目签署备忘录，保障资源的质量，可靠性和可用性，具体定义如下：

可用性 = 站点服务可用的时间 / 总时间

可靠性 = 站点服务可用的时间 / (总时间 - 站点宕机的时间)

每个站点的可用性与可靠性的排名每个月公开发布一次，表现最差的站点将被加入黑名单。这些措施能促使各个站点提高自己的服务水平。

中国数据密集型网格自主开发了一个多功能的网格运行监控系统DIGMOM（如图6）。该系统整合了所有网格相关服务的监控信息。一旦某个组件的服务出现问题，该系统将会通过短信或者邮件的方式向相应的管理员报警，从而保证了网格平台的高效运行。



(A)



图6 整合的网格监控系统：(A) 网络节点的实时状态
(B) 手机短信报警

4. 中国数据密集型网格

4.1 高能物理网格状态

中国一直是国际高能物理合作领域中的活跃成员。来自各个研究所和大学的中国科学家们广泛地参与包括LHC和北京正负电子对撞机BEPCII在内的高能物理实验，因此需要为中国高能物理界建立能够提供强大计算能力的网格计算环境。早在2006年，高能所就与欧洲粒子物理研究中心CERN签订了备忘录，加入国际高能物理网格国际合作，并且建立了数据密集型网格系统并成为国际高能物理网格的二级站点。在中国科学院、国家自然科学基金委和科技部的经费支持下对网格系统进行大规模的建设，到2010年年底，该站点达到1600个CPU的计算能力，640TB的存储能力。为了满足数据传输和作业调度的需求，建立了从国内到欧美的高速网络链接，包括从中国到欧洲之间的1Gbps的带宽以及中美之间的622Mbps的带宽。在中国科学院网络信息中心及清华大学的帮助下对国际网络进行了仔细调试优化，网络链路的延时（影响长距离数据传输性能的重要因素）由350毫秒下降到180毫秒左右，因此整个国际链路的吞吐率得到很大的改进，峰值达到1Gbps以上。图7显示了在中美和中欧数据交换的网络流量24小时统计。

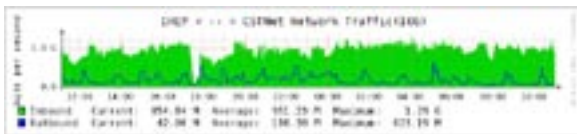


图7 主要网格节点的网络流量图

中国数据密集型网格是包括以高能物理研究所为中心以及卫星站点的联盟。作为中心站点，高能物理研究所站点提供了尽可能多的集中式的网格服务，卫星站点则致力于跟具体项目有关的应用。例

如，山东大学与南京大学站点主要参与到ATLAS实验，中国科技大学的站点最初是为D0实验而建立的，目前已经被转换为一个ATLAS站点。北京大学站点主要是参与CMS实验。图8显示了中国高能物理网格的站点分布。

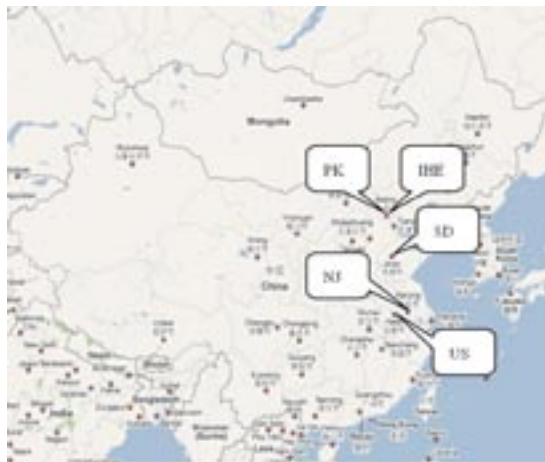


图8 中国高能物理网格站点联合

网格联盟的主要使命是促进各个单位对网格计算环境建设的合作，以及对新的应用的推广，主要活动包括建立各个网格站点之间的高速网络链接，设置网格计算资源，促进网格在高能物理等研究领域的应用。高能物理研究所与法国国家粒子物理与核物理研究院IN2P3，欧洲粒子物理研究中心CERN，以及美国的主要高能物理研究实验室建立了合作关系。通过加入国际高能物理网格，使得中国的高能物理研究人员与世界各地的其他高能物理研究人员能够共享计算与数据资源。

4.2 典型应用

作为数据密集型网格平台的主站点高能所的站点主要用来支持包括ATLAS和CMS在内的高能物理实验，羊八井宇宙线实验等。此外，还对其它学科的科学如生物信息学、地球动力学提供有限度的支持。下面是一些应用的具体例子。

4.2.1 ATLAS

ATLAS^[13]是运行在LHC上最大规模的实验之一，众多研究机构和大学的物理学家们都参与了这项大型的国际合作。作为高能物理网格二级站点，数据密集型网格为中国的物理学家们提供了高性能计算环境，并且为国内的计算任务提供三级站点服务。根据ATLAS计算模型，一个二级站点必须与某一个一级站点相关联，并且该二级站点的任务（如数据传输、作业提交）会取决于与它相关联的一级站点。高能物理研究所选择位于法国里昂的法国粒子物理与核物理研究院计算中心CC-IN2P3作为相关联的一级站点是因为中法两国物理学家之间有着共同的物

理研究兴趣。二级站点需要根据样本数据进行事件重建，这就需要访问一级站点的刻度常数数据库。影响数据库访问效率的一个重要因素是一级站点和二级站点之间的往返延时（RTT），而里昂和北京间的往返延时是212ms。为了提高访问效率，我们测试并且部署了一个Frontier/Squid系统来缓存这些刻度常数。这样，由于减少了远程数据库的访问量，事件处理的速度由原来的每秒0.7个事件提升到每秒11个事件。随后，所有与CC-IN2P3一级站点相关联的二级站点都部署了这个数据库缓存系统，具体细节请参见图9。自去年8月份开始，超过100万个ATLAS作业在高能物理研究所的网格站点上运行，成功率高达88%，这也使得北京站点运行质量名列国际前茅。

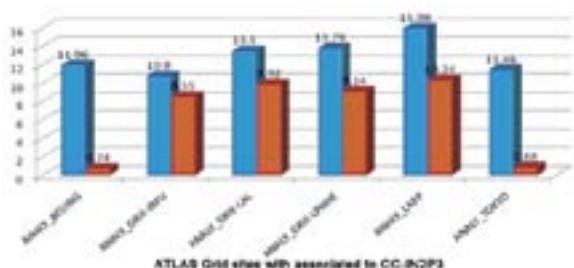


图9 提供数据库缓存服务的时间处理速度，高能所站点的处理速度从0.74个事件/秒提升至 11.96个事件/秒

4.2.2 CMS

高能物理研究所和北京大学是CMS^[14]的参与单位。CMS是类似于ATLAS的一个大型实验项目，但是CMS的计算环境拓扑结构与ATLAS的有所不同。CMS的数据传输拓扑更为灵活，一个二级站点可以与任何一个一级站点进行下载或上传数据，任意两个二级站点之间也可以进行数据传输，这样就要求可以多路由连接到位于不同地域的一级站点。高能物理研究所二级站点与位于美国、欧洲、亚洲的站点建立了连接，每年的数据交换量达到了250TB，完成的作业数超过600,000个。

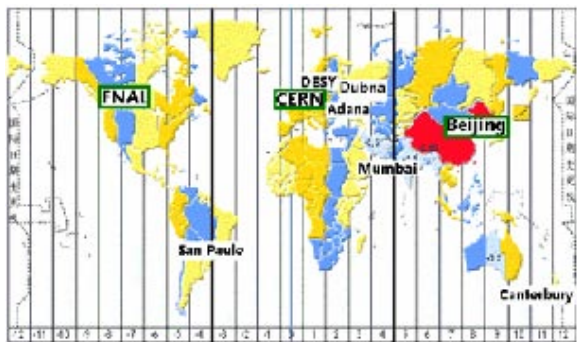


图10 CMS区域性运营中心

高能物理网格共享的不仅仅是计算资源，还有人力资源。高能物理研究所于2009年6月为CMS的计算网格建立了名为CMSROC@Beijing的区域运营中心。CMSROC@Beijing是继美国的Fermilab和德国

的DESY之后的第三个区域运营中心。这是CMS首次将远程运营从欧洲、北美扩展到了亚洲。如图10所示，三个运营中心分别位于三个不同的时区，每个运营中心值班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，区域运营中心也可以帮助中国的物理学家们更方便地参与CMS实验的研究活动。

4.2.3 蛋白质结构预测

数据密集型网格还对除高能物理以外的其他学科的应用提供支持。通过与罗马第三大学及雅盖隆大学医学院的合作，我们搭建了一个用于预测海量蛋白质结构的模拟环境。生物学领域有一个被称为“未诞生的蛋白质”问题，非天然的蛋白质理论上要比天然的蛋白质数量多得多，而我们的想法是能否通过计算预测来探讨是否存在非天然蛋白质能够被折叠成稳定结构并表现出与天然蛋白质相似的功能特性甚至生物化学特性。这些计算任务由中国、意大利、波兰和希腊的网格站点共同承担。对大量非天然蛋白质序列结构的预测显示大部分的蛋白质序列都能形成稳定的三维结构。图11展示的是用于蛋白质结构预测的可视化工具。

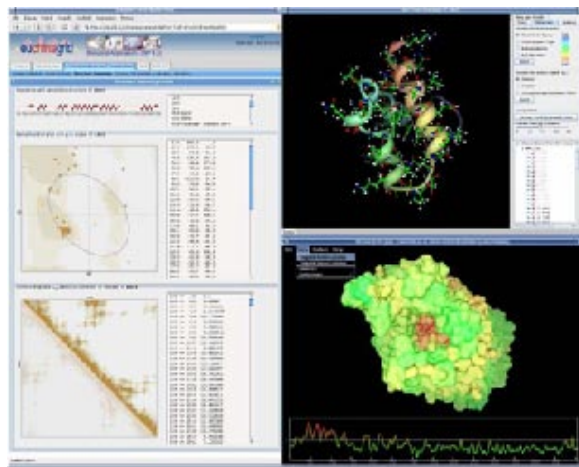


图11 从未诞生过的蛋白质预测

4.2.4 蛋白质折叠

网格平台还支持了生物物理研究。大连化学物理研究所利用网格平台进行了具有高度序列一致性但不同三维结构的蛋白折叠机理的理论研究。这个课题是理论研究两个具有高度序列一致性的蛋白质却具有不同三维空间结构的微观机理。在生物物理理论里有如下假设，即两个蛋白质如果序列高度一致，他们三维空间结构也具有很高的相似性。但是在人工设计的两个蛋白质里我们看到了反例，GA和GB，他们最原始的序列相似性只有~30%，空间结构决然不同，但人们把他们的序列突变到具有88%和95%等同性的时候，他们的三维空间结构已然不同，基本保持着原来的样子。

这是一个非常有意义的例子，对于认知蛋白质折叠的机理，序列、结构和功能之间的关系具有非常重要的影响。大连化学物理研究所进行了大规模的、一系列不同温度下的分子动力学模拟，希望找到它们折叠机理的不同和关键之处。通过计算找到了它们折叠的过渡态，发现了主导它们各自折叠规律的重要相互作用，为进一步人工从头设计和改造蛋白质奠定了一定的理论基础（如图12）。

5. 结束语

中国数据密集型网格已经为中国高能物理学界，尤其是LHC实验准备好了科学计算基础设施。高能物理网格将世界上不同领域不同地域的计算资源

和科学家聚集到一起，实现高能物理研究的协同与资源共享。高能物理网格的研究和建设同时推动了计算机与网络等相关技术的发展并吸引更多领域加入高能物理网格，使其成为一个综合性的科学计算平台。计算和存储资源会随着参与站点的增多而增加，更多学科而不仅仅是高能物理也会参与进来。

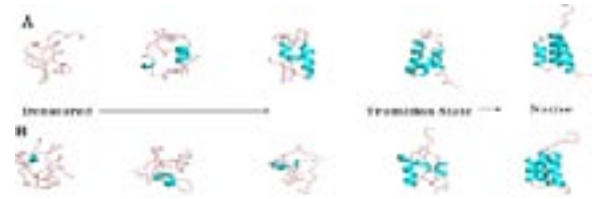


图12 高度序列一致但不同三维结构的蛋白折叠过程

参考文献：

- [1] LHC site: <http://public.web.cern.ch/public/en/LHC/LHC-en.html>
- [2] BEPCII: <http://www.ihep.ac.cn/BEPCII/index.html>
- [3] BESIII: <http://bes3.ihep.ac.cn/>
- [4] WLCG site: <http://lcg.web.cern.ch/LCG/>
- [5] The Enabling Grids for E - sciencE (EGEE) project: <http://public.eu-egee.org/>.
- [6] The Open Science Grid site: <http://www.opensciencegrid.org/>.
- [7] Globus: <http://www.globus.org/>
- [8] gLite: <http://glite.web.cern.ch/glite/>
- [9] GOS: http://www.cngrid.org/download/CNGrid_GOS/CNGrid_GOS_zh_CN.html
- [10] Naregi: http://www.naregi.org/index_e.html
- [11] IGTF: <http://www.igtf.net/>
- [12] CNGrid中国国家网格: <http://www.cngrid.org/web/guest/home>
- [13] ATLAS实验: <http://atlas.ch/>
- [14] CMS实验: <http://cms.web.cern.ch/cms/index.html>