北京谱仪软件、数据处理与网络的开拓

许榕生

（中国科学院高能物理研究所 100049）

我是1978年高能所第一批硕士研究生，1982年自费公派到美国继续攻读高能物理博士学位，1988年应李政道先生推荐，返回高能所参加北京正负电子对撞机实验工作。高能所以“急需人才”的名义安排了我，在高能所主要经历了三件事，即北京谱仪软件的完成、建立数据处理规范和开通国际互联网专线等。谨借高能所建所50周年之际，撰写回顾献作纪念。

一、北京谱仪的软件

作为高能物理对撞机实验的离线分析基础软件主要包括实验数据分析与Monte Carlo模拟两个基本部分。因为北京正负电子对撞机及北京谱仪的设计与美国斯坦福直线加速器中心（SLAC）的MarkIII探测器及实验环境最为相似，所以，北京谱仪离线分析的基础软件自然选定以SLAC的MarkIII组进行参考并进行移植，再加之相应的修改和补充完善，运行时还加上西欧中心（CERN）的CERNLab以及美国几家实验室提供的软件库。

1988年北京谱仪开始获取实验数据，急需离线分析软件来实现对数据的处理和进一步物理分析，我的第一个任务就是协助完善北京谱仪的软件工程，建立实验数据处理的操作规范。我曾在美国MarkIII组七年时间，对他们的离线分析软件熟悉，了解到该软件也是从MarkII和MarkI延续下来的，由众多的实验物理学家和研究生用十多年时间编写积累而成，该软件以FORTRAN语言为主，运行在IBM大型机上。当时在高能所只有DEC计算机环境，FORTRAN程序可以通用，但除了程序中必须对实验环境的参数修改外，由于离线分析程序属I/O密集型软件，即需要不断地读写存储设备上的事例数据进行相应的运算，因而在软件中与外部存储设备打交道的配置参数也必须重新调整。此外，离线分析运行的用户操作计算机的接口程序（User Interface）我们必须进行自行的设计与编写，这是离线分析一系列工作的基本保证。

当时，软件组的成员对基础软件（FORTRAN源程序）部分的审核和改造已经做了大量工作，尤其Monte Carlo事例模拟程序已先期基本完善。然而，前面提到的这个接口程序是要自己动手由DEC操作系统的指令语言来编写，不是一般用户的高级语言编程。我用了不长的一段时间掌握了DEC机器的底层语言，编写出了北京谱仪离线分析统一的用户接口（DRUNK），它适用于任何需要从存储中读写数据进行事例显示、数据处理、事例统计以及物理分析等全套离线分析的功能；同时也编写了事例模拟的用户接口（SOBER），它适用于产生各种模拟数据并录入相关的存储设备。这两个用户接口系统完成后，北京谱仪的离线分析可以大踏步地迈进了，首先是实验数据的事例过滤，其次是事例分类、事例重建以及物理分析等的各个软件都可以实际进行调试和运行。

这时谱仪收集到的第一批实际数据开始将数据磁带挂上大型计算机运行相应的统计程序来观察实验的最终效果。当北京谱仪积累到约300万左右的事例时，我用MarkIII组的相关程序检验了一下，看出来我们的数据是正常的，软件程序框架也是可用的，我心中有了数并告诉大家这一见好的迹象，包括J/Psi粒子衰变道中重要的Dalitz图的显示等。于是接下来对北京谱仪获取的数据进行了日夜不停的过滤、分类和事例重建（物理描述），进一步看出了探测器各部位数据的质量以及初步的物理信号，物理分析的结果指日可待！北京谱仪软件系统的关键作用受到了好评，我很荣幸地被列入北京正负电子对撞机荣获国家科技进步特等奖的集体名单中（1991年），个人也被中国物理学会授予了胡刚复奖（1993年）。

二、北京谱仪的数据处理概述

作为统计处理与物理分析的原始实验数据(Raw data)通常是不完整的,含有噪音、数据缺失和错误等情况出现。这种存在缺陷的“脏”数据,不能直接进行数据统计和物理分析,如果直接进行分析,其结果必然差强人意。为了提高数据的质量、减少不必要的数据处理开销,就必需对这些数据进行一些预处理,即数据过滤和数据刻度。当今社会上的大数据分析称之为数据清理和数据校准。

数据预处理的过程就是负责将分散的、异构数据源中的数据处理成精致数据。数据预处理是整个数据统计、数据挖掘过程中很重要的一个步骤。实际工作中,数据预处理花费的工作量约占整个数据分析近半以上的工作量。当数据成为“干净数据”、甚至认为是高质量的数据后,对后续的离线分析将产生重要的积极的作用。反之,如果对数据预处理的支持力度不足,研究不够,这与其的重要地位是很不相称的。为此北京谱仪研究室专门成立了一个数据组和刻度组(我是首任数据组组长)来从事这项繁琐而重要的工作。

根据经验或统计假设检验,我们首先观察到在线获取的实验数据含有各种意外电信号干扰记录下来的很多事例,这种事例的数据块长度比正常的事例大出好几十倍,在数据磁带上占据了大量空间。如果不把这些噪声数据(约占1/3)清除出去,势必造成计算机资源的极大浪费,包括I/O的消耗等。根据事例数据块大小的统计分析,给定噪声异常数据的删除标准,有效过滤了垃圾数据、大大节省了存储空间和后面阶段的数据分析时间。当年数据组的工作是非常辛苦的,因为计算机作业24小时运行过程中必须用人工装卸磁带,所以要安排人员值夜班。此外,所有过滤以后的数据磁带和分类后的数据磁带都要贴上标签、注明标号,要求精细操作以保证后续离线工作的正确及方便。如果把后续的物理分析比作高级大厨烹调美味佳肴,那么数据预处理就好比把所有的菜料加工的一道必要工序。

实际上对过滤异常数据的处理方法可以进一步研究引进许多统计方法,例如用3σ法则、数字滤波(最小二乘滤波、维纳滤波及卡尔曼滤波)以及人工智能等处理噪声的工具都可以做深入探讨。因高能所需要建立国际计算机联网专线,我被调往高能所计算中心去完成新的使命,数据组的工作留给了其他人员继续。此外,与数据预处理密切相关的事例重建工作由刻度组进行,重建好的数据存储为DST数据提供给物理分析使用。总之,北京谱仪数据组与刻度组都是离线分析的重要环节,北京正负电子对撞机工程培养出了国内最早的一批数据工程师,包括具备既了解高能物理实验又熟悉数据分析的交叉专业人才(即今天大数据分析中的CDO,Chief Data Officer)。

三、建立中国第一条互联网专线及WWW服务器

高能所率先在国内开通互联网国际专线是北京正负电子对撞机工程发展的需要,1991年6月1日美国潘诺夫斯基教授(中科院外籍院士)写给中共中央常委宋平和科学院院长周光召的报告里提出需要建立一条中美之间高速的电子数据网线以保证北京谱仪BES国际合作组的实验数据远程传输需要(图1)。之前高能所也曾多次尝试与欧洲开通计算机联网,但由于不具备高速联网的条件,基本上只能进行电子邮件的尝试。进入九十年代Internet(互联网)的诞生并应用到高能物理研究上,24小时运行的TCP/IP协议在各种计算机网络间互联互通,除了用它进行海量实验数据的远程传输外,同时实现包括电子邮件、远程登录等多项重要的功能。

面临新的机遇,高能所遇到的挑战在于国内没有任何现成的经验可以借鉴,而电信部门还没有提供互联网的接入服务。在BES国际合作组的网络专家建议下,高能所及时订购了CISCO路由器,并随即前往北京电信局申请带宽64K的专线。由于当年还没有光纤通到高能所,于是电信部门临时采用了两根电话线代替光纤的办法,在国际段采用卫星连接与跨洋的美国计算机网络实现互联。高能所计算中心协助电信局经过18个月线路的反复调试,于1993年的3月这条链路的通信误码率降低到规定的范围,高能所终于开通了国内第一条Internet专线。一年多后高能所改卫星线路为海底光缆并扩容到128K,从日本高能所KEK直接联入Internet网。

尽管高能所这条专线的速率不算很高,但是立

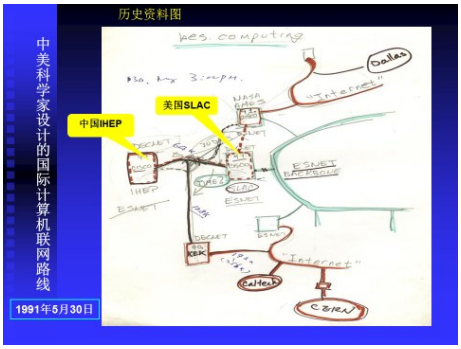

图1　1991年在高能所(IHEP)绘下的一张国际计算机联网设计图,第一步接通斯坦福加速器中心(SLAC)接入美国能源部网络(ESnet)。第二步通过日本高能所(KEK)直接联入国际互联网

即显示出了效果。北京谱仪的软件升级,几篇文章的发表都得益于依靠互联网的快速讨论和修改,尤其是高能所利用镜像代理技术对在北京举办的国际高能物理会议(1995年)实现了远程视频的转播,令国外与会者都感到欣慰。

那一年我到国家自然科学基金委介绍Internet时提出可以协助他们建立电子邮件,基金委领导当场决定拨款委托高能所为全国几百个课题组负责人开通电子邮箱,于是高能所增加了直拨电话线,并根据基金委提供的专家名单开户用电话拨号上网。高能所计算中心认真负责地为每一位用户演示安装;技术员安德海专门编写了一个很有创意的用户接口界面,让大家事先把邮件的内容、地址写好,一旦拨号成功即自动完成接收与发送用户邮件。这样高能所的电子邮箱系统节省了这些科学家大量的时间和费用,大大方便了他们与世界各地的科技交流。正是这批重要的科学家率先尝到了互联网的威力,不久他们以学部委员会的名义提出了尽快建立国家级的互联网,也就是后来的国家科技网(CSTNet)和全国教育网(CERNet)。

影响全球信息技术发展的WWW技术是在西欧核子研究中心(CERN)发明的,1994年初高能所批准我到日内瓦的西欧中心访问,遇见了WWW的发明者蒂姆.B.李(Tim Berners-Lee),在他的办公室里我看到WWW技术的重要历史作用。当年的4月15日我组织高能所建立了中国第一台WWW服务器,域名是http://www.ihep.ac.cn/,沿用至今。这项技术极大地推动了国内信息化的发展,当时亚洲地区也刚刚出现网站。高能所这台WWW服务器(图2左)于2021年8月作为科技文物被北京国家博物馆收藏。当年计算中心的研究生樊岚设计了高能所第一套英文版的网页(图2右)。

高能所开通互联网、建立第一台WWW对中国进入互联网时代做出了突出贡献,而后的网络安全课题研究又为抗拒网络攻击以及抵御黑客犯罪、加强我国互联网安全意识等方面不断做出了显著的成果。其中我和高能所计算中心同仁承担的网络安全课题研究获得中国科学院科技进步一等奖(1999年),及国家科技进步二等奖(2001年)。高能所还为国家培养出了一批最早的网络安全领军人才,他们正在国家各个部门及重要企业发挥着重大的作用。

四、结语

高能所五十年来成果辉煌,一代又一代杰出的科学家与科技人才辈出。我有幸在高能所几十年学到了许多知识、也出了些许绵薄之力,由衷地感谢高能所的领导、导师和同事们共同创造的环境和机遇!让我们衷心地祝愿高能所未来更加美好。

图2　1994年高能所建立的国内第一台WWW服务器(左)及第一张高能所的网页(右)