王彬,肖文名,李永生,等. 华南区域中心计算资源管理系统的建立与应用[J]. 气象,2011,37(6):764-770.

华南区域中心计算资源管理系统的建立与应用*

王 彬[1] 肖文名[2] 李永生[2] 常 飚[1] 陈晓宇[2]

1 国家气象信息中心,北京 100081

2 广东省气象信息中心,广州 510080

提 要:近几年华南区域气象中心高性能计算机系统能力建设迅速,总能力超过 2 万亿次。为更好地发挥区域内计算资源投资效益,在引进吸收中国气象局国家级高性能计算资源管理软件的基础上,结合区域实际需要进行设计与自主开发,成功建立了一套精细化计算资源管理软件。该系统采用了"资源账户-资源点数"的设计思想,针对不同类型用户提供了相应的用户组织方案。该系统现已完成了编码实现和部署,实现了区域内两个主要高性能计算机系统的资源管理功能。应用到业务后,发挥了良好效益,实现了资源的均衡充分利用,成为区域内计算机资源管理的得力工具。

关键词:计算资源管理,高性能计算机系统,华南区域气象中心,资源点数,资源账户

Establishment and Application of South China Regional Meteorological Centre Computing Resource Management System

WANG Bin[1] XIAO Wenming[2] LI Yongsheng[2] CHANG Biao[1] CHEN Xiaoyu[2]

1 National Meteorological Information Centre, Beijing 100081

2 Guangdong Meteorological Information Centre, Guangzhou 510080

Abstract: In the past few years, rapid progress has been made in high performance computer capability building in South China Regional Meteorological Centre. The total capability exceeds 2 TFLOPS in theoretical peak performance now. In order to make the most of the computing resources invested, a fine-grained computing resource management system is successfully designed and developed, which draws much from national meteorological high performance computer resource management software and takes into consideration the actual demands of South China Regional Meteorological Centre. The design idea of "resource accounts" and "resource credits" is adopted by the resource management system and thus, different kinds of users are provided with corresponding user organization solutions. Code implementation and deployment have been done with two leading high performance computer systems under management in South China Regional Meteorological Centre. Since it has been applied into operations, the resource management system produces satisfactory effects, implements balanced and full use of resources and becomes the right hand for system administration technicians.

Key words: computing resource management, high performance computer system, South China Regional Meteorological Centre, resource credits, resource account

* 公益性行业(气象)科研专项"基于网格的气象计算资源管理技术研究"(GYHY200806018)和气象关键技术集成与应用项目"计算资源管理共享技术应用集成与区域推广"(CMAGJ2011M66)共同资助

2010 年 4 月 28 日收稿; 2010 年 12 月 19 日收修定稿

第一作者:王彬,主要从事气象高性能计算、网格计算业务与开发研究工作. Email:wangbin@cma.gov.cn

引　言

华南区域气象中心是中国气象局八大区域气象中心之一,覆盖了广东、广西、海南等三个省区。华南区域气象中心地理位置居于欧亚大陆南端,北依南岭,南濒南海,受低纬度热带天气系统和中高纬度天气系统的交替影响,天气气候复杂多变,是我国灾种多、频率高、危害重的灾害大区。

高性能计算机(High Performance Computing,HPC)系统是支撑天气预报与气候研究的重要技术平台。在中国气象局和区域内地方政府的大力支持下,"十一五"期间华南区域气象中心高性能计算机能力建设迅速。现已引进了两套高性能计算机系统,能力均超过万亿次,合计达 2.15 万亿次。

取得能力建设大发展之后,如何把巨资购买的计算资源管理好、利用好,最大化地发挥资源投资效益就成为区域中心计算机管理最重要的任务。而区域中心内高性能计算资源管理水平的落后和业务系统软件的缺乏大大制约了对资源使用的了解,也不能实施有力的分配控制。

为此,华南区域中心在利用气象计算网格技术[1-6]整合区域内高性能计算机资源基础上,引进吸收了中国气象局国家级高性能计算资源管理软件[7-8],结合区域实际需要进行设计与自主开发,成功建立了一个区域内跨地域的、分布式的精细化计算资源管理软件,并投入到稳定业务运行中。

1　资源管理系统设计

1.1　设计思想

资源管理系统的设计目标是通过提供强大的技术手段,帮助实现高性能计算机资源合理和高效的利用[9-11]。资源管理可分为资源使用记账、资源统计分析和资源分配管理等主要功能。这三个功能互相依赖,互为前提。资源的合理高效利用的前提就是对用户进行合理高效的资源分配,而制定资源分配策略依赖于对资源使用的精准掌握和分析,资源分析统计又需要精细粒度的资源使用记账数据。

系统设计思想借鉴了银行的信用卡账户管理,采用"资源账户-资源点数"[12]作为整个系统的核心机制,如图 1 所示。

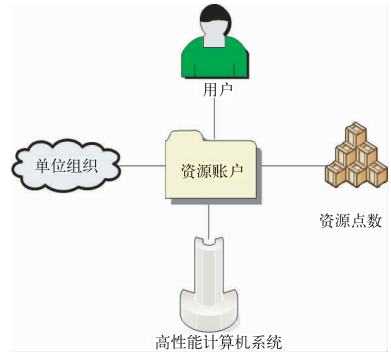

图 1　"资源账户-资源点数"设计思想

Fig. 1　Design idea of "resource accounts" and "resource credits"

资源账户(resource account)是实施资源管理的基本单元,可以按照高性能计算机用户的组织机构和形式建立若干个账户,账户之间可以存在一定的层次关系。资源账户包含了用户、单位组织、计算机系统和资源点数等方面的要素。

资源点数(resource credits)是实现资源管理量化的关键。通过将抽象的计算资源与时间的组合换算成具体的资源点数,从而可以将计算资源在一段时间内的使用量进行定量精确的分配,并定量的记录下用户对计算资源的占用情况,了解资源的使用情况。

一个资源账户既可以属于某一个具体的用户,又可以属于某一个具体的单位组织或项目。通过信用卡常用操作方法,可以方便地实现计算资源的各种管理功能。

(1) 资源分配管理:为用户(项目)的资源账户注入资源点数就完成了计算资源的分配,可以注入不同数量和使用期限的资源点数,从而实现了不同的资源分配策略。

(2) 资源使用记账:一次作业完成后,将根据占用的资源量和时间从作业提交者的资源账户中扣除相应数量的资源点数,从而完成资源记账。

(3) 资源统计分析:查询资源账户内的资源点数的使用和余额情况,并进行比较排序求和等运算,即可实现资源统计与分析。

1.2　用户组织方案

根据华南区域气象中心用户组织的实际情况,基于"资源账户-资源点数"思想,设计了资源用户组织方案。

根据高性能计算机用户用途的不同,分为业务和科研两种类型(见图 2)。根据不同类型用户的特点分级管理。

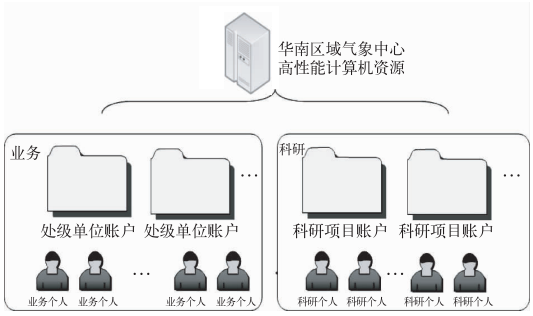

图 2 高性能计算机用户分级管理

Fig. 2 Hierarchical management of HPC users

(1) 业务

业务用户使用资源的时间和数量相对固定,与本单位的业务职责密切相关,设置业务用户个人和所属处级单位两个层次。

(2) 科研

科研型用户使用资源的时间和数量具有随机、临时组合的特点,一般随着项目课题的实施进行。为此,设置科研项目和科研用户个人两个层次。

1.3 资源度量与分配单位"GCU"

需要引入一种合理的计量单位来描述计算资源,使计算资源使用量得以量化,量化后的资源使用量可以精确地描述用户和应用所使用的计算资源,计算各种应用所占用资源的比例。在此基础上,使计算资源的分配、使用记账和统计分析等管理功能都能以定量的方式实现。

借鉴了国家级高性能计算机资源管理思路,引入了虚拟计算资源单位——(General Computing Unit, GCU)作为资源度量与分配管理的单位[7]。1 个 GCU 相当于高性能计算机系统 1 个 CPU 小时的计算能力。通过一个统一的计算单元"GCU",实现了各种不同架构、不同型号计算机系统计算资源的统一计量。

1.4 功能模块设计

从整体上看,华南区域中心计算资源管理系统可分为计算资源层、资源管理层和用户接口层等三个层次,如图 3 所示。

(1) 计算资源层

计算资源层由华南区域气象中心内分布各地的高性能计算机系统组成。计算资源可以为气象网络中的任意节点,分布在区域中任意物理位置。目前主要的高性能计算机系统分布在广州和东莞两地,

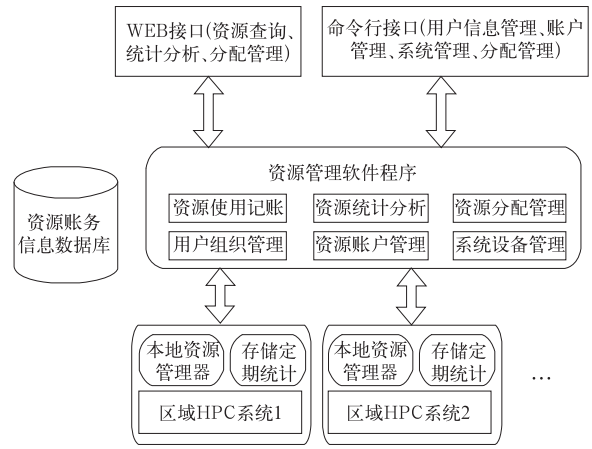

图 3 资源管理系统功能模块层次结构

Fig. 3 Function layers and modules of resource management system

各地区的系统通过本地资源管理器实时记录资源管理信息并发送到资源管理层。

高性能计算机系统的本地资源管理器一般是指本地的作业管理系统,如 LoadLeveler、PBS(Portable Batch System,便携批处理系统)、LSF(Load Sharing Facility,负载共享工具)等。存储定期统计模块实现了本地存储资源的定时记录与统计。

资源管理层主要通过与本地资源管理器和存储定时统计模块的直接通信交互,或者是通过网格平台,实现各种资源管理功能。

(2) 资源管理层

资源管理层由资源管理软件程序和资源账务信息数据库组成。

资源账务信息数据库通过数据表的形式记录了资源账户、用户组织、计算机系统、资源分配和作业记账等信息。

资源管理软件程序是整个系统的核心,包括资源使用记账、资源统计分析、资源分配管理、资源账户管理、系统设备管理和用户组织管理等功能模块。

• 资源使用记账:实现了计算机用户提交的作业在运行过程中使用资源信息的记账入库。记账过程基于用户提交的每一次作业,因此资源管理达到了最细的粒度,为资源使用的统计分析奠定了基础。

• 资源统计分析:从不同的维度(时间区间、用户、单位或项目、计算机系统等)查询资源使用量,即可获取丰富的可定制的统计信息,如资源点数使用量统计、运行作业数统计等。基于统计信息可生成不同的分析结果,来指导资源的分配和调度策略。

• 资源分配管理:以 GCU 为计算单元,实现了资源的预分配、扣除、透支、计算等分配管理功能。

• 资源账户管理:实现了资源账户的创建、删除、修改、锁定/解锁等功能,支持资源账户的组/用户之间的嵌套树状关系管理。

• 系统设备管理:实现了高性能计算机系统信息的增加、删除与修改等管理功能。

• 用户组织管理:实现了用户信息、单位部门和项目信息的增加、删除与修改等管理功能。

(3) 用户接口层

最上层为用户接口层,是用户使用系统的访问接口,包括 WEB 和命令行两种形式。WEB 接口包括资源使用量查询、资源统计与分析等功能。命令行接口提供了系统的全部功能,包括用户管理、账户管理、系统管理、资源分配管理、资源记账、资源统计分析与报表等功能。

2　系统实现

2.1　系统运行流程

系统的运行流程如图 4 所示。

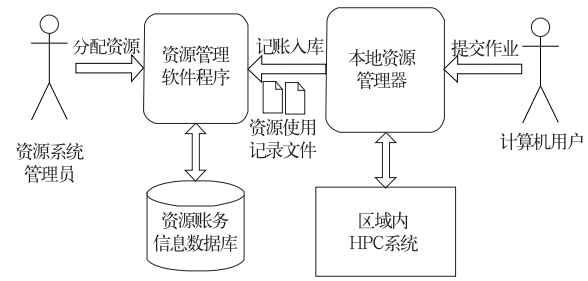

图 4　系统运行流程图

Fig. 4　System runtime procedures

运行过程可分为 7 步:

(1) 资源系统管理员通过向资源账户存入资源点数为计算机用户分配资源;

(2) 计算机用户提交作业给某个集群系统的本地资源管理器;

(3) 获取资源报价;

(4) 资源预扣除;

(5) 启动作业;

(6) 完成作业;

(7) 根据作业占用的计算资源和时间,扣除资源账户相应的资源点数。

2.2　系统实现与部署情况

按照系统的设计方案,成功地在华南区域气象中心内完成了软件模块代码开发和软件配置部署工作。

资源管理软件完成采用实用报表提取语言(Practical Extraction and Report Language,PERL)技术开发,资源账务信息数据库采用了开源的数据库[13]技术实现。

• 资源记账模块:每个作业运行完后,本地资源管理器生成作业运行日志,资源记账模块读取日志,根据日志内容更新数据库,完成资源记账和作业信息记录。

• 用户信息、组织信息、计算机系统信息、资源账户和资源分配等管理模块:使用命令封装数据库操作,实现各类信息的增删改等功能。

• 资源统计模块:使用命令封装数据库查询操作,实现任意时间段内按照用户、组织和计算机系统多方位的统计。

• 报表模块:在统计模块基础上实现按月生成资源统计月度报表。

报表模块每月生成的资源统计月报定时上传到 Web 服务器,Web 应用程序以月报为数据源,通过图形模块根据月报生成各类图形文件,在页面中以图形和表格的形式显示资源统计信息。

系统实现部署情况如图 5 所示。

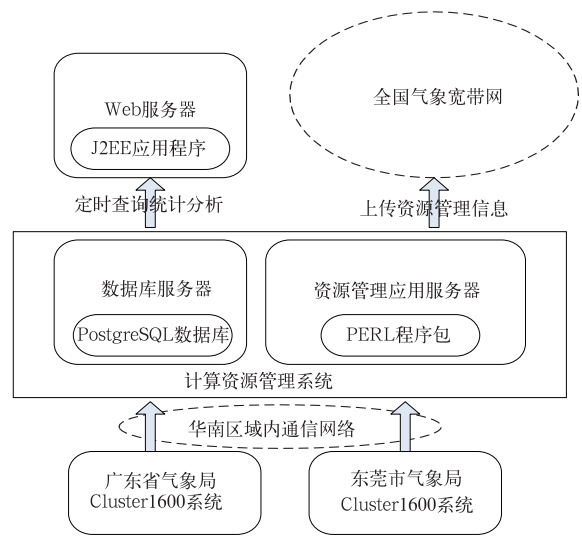

图 5　系统实现部署情况

Fig. 5　System implementation and deployment

资源管理软件打包部署在资源管理应用服务器上运行。资源账务信息数据库运行于一台独立专用数据库服务器上。

现已在华南区域气象中心的广东省气象局 IBM Cluster1600 系统(位于广州)、东莞市气象局 IBM

Cluster1600 系统(位于东莞)上部署运行了资源管理守护进程,通过配置系统上的 LoadLeveler 的前导和后续脚本,嵌入到本地作业管理流程中。通过区域内通信网络运行交互,实现了对这两台位于不同地理位置的高性能计算机系统的精细化资源管理。

按照国家气象计算网格资源管理与监视的统一要求,从数据库定期查询统计,生成区域内资源信息文件,经全国气象宽带网自动定时上传到中国气象局[14-15]。

2.3 资源分配策略与方案

通过统一的量化手段记录资源的数量,以存入资源点数的形式对资源用户进行一定时间区间内资源的分配,进而精确的记录、分配和控制用户资源使用量。以计算资源使用的最小量作为资源记录的单位,实现最细粒度的资源记账和分配控制。

高性能计算机系统各种资源优先满足业务类型用户,并提供充足的备份资源。在此前提下,尽可能满足气象科研与开发的需要。

对于区域内单个高性能计算机系统,每年可用的高性能计算机计算资源总量(HPC Resource Credits,HPCRC)为:

$HPCRC = CPU$(核)个数 $\times$ 24 小时 $\times$ 365 天

则区域内每年可用的计算资源总量 $HPCRCA$ 为:

$$HPCRCA = \sum_{i=1}^{n} HPCRC_i$$

其中 n 代表华南区域内高性能计算机系统总数。

对于业务型用户计算资源的分配,根据应用的评估结果,确认用户所需要的资源量。以年为单位,分配用户在这一年内的总的资源量。在年末对用户的使用情况进行评估,在下一年重新分配相应的资源量。

对于科研型用户资源的分配,则根据承担项目课题的工作量和实施周期进行分配,资源分配点数具有生命周期时效性,一旦项目完成,资源分配量则自动到期取消。

3 系统业务应用

计算资源管理系统投入业务运行以来,业务保障措施到位,系统运行基本平稳。作为日常业务的一部分给予保障,制定了相关规章制度。专人负责运行维护,清晰掌握业务流程,详细记录故障处理过程。在计算资源管理系统的支持下可以统计不同时间尺度下的资源使用量和资源利用率,为系统管理员和用户掌握系统资源实际使用情况提供了可靠的数据支持,并在此基础上对资源使用情况进行综合分析,制定更加合理的资源使用策略,从而更好地发挥资源的效益。

3.1 资源统计与分析

(1) 全年逐月资源统计与分析

以广东省气象局 IBM Cluster1600 系统为例,对全年资源的使用情况和系统利用率进行了统计分析,如图 6 所示。

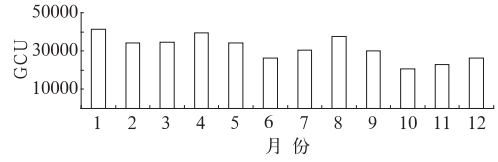

图 6 广东省局 IBM Cluster1600 系统 2009 年每月资源使用量

Fig. 6 Monthly resource usage of Guangdong IBM Cluster1600 in 2009

由图 6 可知,系统每月资源的使用量基本稳定在 40000GCU 左右,分析可知业务用户和科研用户的资源使用情况基本稳定。

(2) 逐日资源统计与分析

通过计算资源管理系统查询得到的 2009 年 8 月华南区域气象中心两个高性能计算机系统逐日平均利用率和资源使用量。

东莞市气象局 IBM Cluster1600 系统逐日平均利用率如图 7 所示。

(3) 逐小时资源统计与分析

利用计算资源管理系统,可获取精细到小时的资源使用量,进而可以分析系统负载的波峰和波谷,判断资源使用趋势。这里随机选取了 2009 年 10、11 和 12 月中 1 天广东省气象局 Cluster1600 系统每小时资源使用情况作为分析对象,如图 8 所示。

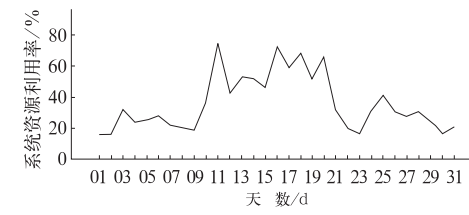

图 7 东莞市局 Cluster1600 系统 2009 年 8 月逐日系统利用率

Fig. 7 Daily resource usage of Dongguan IBM Cluster1600 in August 2009

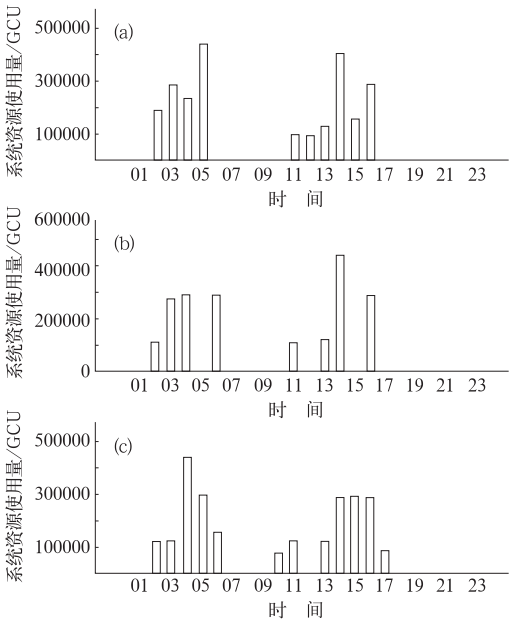

图 8　2009 年 10(a)、11(b)和 12 月(c)广东省局 Cluster1600 单日逐小时资源使用量

Fig. 8　Hourly resource usage of Guangdong BM Cluster1600 in October (a), November (b) and December (c) of 2009

对单日 24 小时的使用情况分析可知,广东省气象局 Cluster1600 系统的使用量呈现“两个峰值”的特点,分别是每天的 02—06 时和 11—16 时。这也为系统管理人员进一步作业调度提供了可靠的依据,在以后系统使用过程中指导用户尽量利用系统的空余时间,从而能够更加有效地利用系统的计算资源,提高系统的利用率。

3.2　资源调度应用分析

在对华南区域中心计算资源使用情况进行全年、逐月和逐日详细统计的基础上,对资源使用情况进行了系统的分析和梳理,并对计算资源用户提出了具体的使用建议。

经过一段时间的实际业务运行,系统资源使用效率有了明显改善。经过调整,目前系统资源基本满足主要业务数值预报模式实现备份运行的资源需求,在计算资源管理系统的支持下,在原有业务模式系统没有重大变更的情况下,合理安排调度资源满足了新增亚洲运动会数值预报模式的计算资源需求,并实现模式的备份运行。调整以后,华南区域两台主要高性能计算机的负载更加均衡(如图 9 和图 10 所示),一方面分散了业务系统的运行故障风险,另一方面也为新增数值预报业务计算需求与原有业务计算需求调整提供了一定的调整空间。

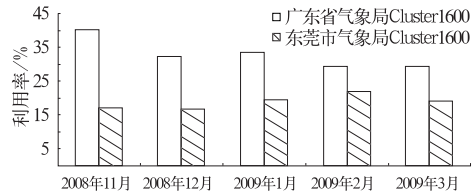

图 9　调度调整前高性能计算机利用率对比

Fig. 9　Comparison of resource usage before resource scheduling and adjustment

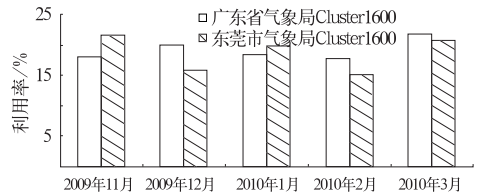

图 10　调度调整后高性能计算机利用率对比图

Fig. 10　Comparison of resource usage after resource scheduling and adjustment

4　结　语

华南区域中心计算资源管理系统投入业务运行以来,发挥了良好的效益。系统管理员能够利用该系统实时、精细、方便地查询、统计和分析各类型用户及所属组织使用计算资源情况,统计分析结果为计算机系统资源的合理分配、业务备份和均衡使用提供了重要参考依据。该系统已成为华南区域中心计算机资源管理的得力工具。

计算资源管理系统未来将推广应用到区域内的深圳市气象局、广西、海南省气象局,进一步提升华南区域气象中心高性能计算网格平台资源的使用效益。

参考文献

[1]　王彬,宗翔,田浩. 国家气象计算网格的设计与建立[J]. 应用气象学报,2010,21(5):632-640.

[2]　李永生,王彬,肖文名,等. 广州区域中心气象计算网格节点的设计与实现. //国家气象信息中心 2008 年度科技年会论文集[C],北京,2009:30-36.

[3]　杨学胜,张卫民,陈德辉. 网格计算及其在气象中的应用[J]. 气象,2005,31(2):79-82.

[4]　王彬,周斌,魏敏. 气象计算网格模式预报系统的建立与优化[J]. 计算机应用研究,2010,27(11): 4182-4184.

[5]　马廷淮,闫荞荞,王彬,等. 基于元任务的网格调度算法综述[J]. 武汉理工大学学报,2010, 32(16):143-147.

[6] 赵威，李明皓，唐远明，等. 辽宁省气象网络计算应用系统的设计与实现[J]. 气象，2009，35(12)：133-138.
[7] 王彬，宗翔，魏敏. 一个精细粒度实时计算资源管理系统[J]. 应用气象学报，2008，19(4)：507-511.
[8] 宗翔，王彬. 国家级气象高性能计算机管理与应用网络平台设计[J]. 应用气象学报，2006，17(5)：629-634.
[9] 王英，常骏，李永利，等. 内蒙古气象综合信息系统功能设计与实现方式[J]. 气象，2010，36(4)：80-84.
[10] 王玉彬，周勇，梁丰，等. 2008 年北京奥运会气象服务中的信息资源整合[J]. 气象，2009，35(6)：109-117.
[11] 胡胜，罗兵，黄晓梅，等. 临近预报系统(SWIFT)中风暴产品的设计及应用[J]. 气象，2010，36(1)：54-58.
[12] Jackson S. Allocation Management Solutions for High Performance Computing. Proceedings of PDPTA 2005[C]，Athens：CSREA Press，2006：10-16.
[13] PostgreSQL 8.3.9 Documentation[EB/OL]. http：//www.postgresql.org/files/documentation/pdf/8.3/postgresql-8.3-A4.pdf.
[14] 王彬，常飚，朱江，等. 气象计算网格平台资源监视模块的设计与实现[J]. 应用气象学报，2009，20(5)：642-648.
[15] 李湘. 气象通信系统发展与展望[J]. 气象，2010，36(7)：56-61.

新书架

周秀骥文选

周秀骥 等著

周秀骥院士是我国著名的大气科学专家，在推动、发展和开拓中国大气物理学等方面做出了卓越贡献。他治学严谨，探索创新，提出独到的学术见解，形成自己的学术观点，为我国大气科学和气象事业的发展做出了开创性的贡献。本书收录了他从事气象事业以来的代表性论著，主要涵盖云雾物理和雷电物理、大气遥感、中尺度气象学、环境与气候和大气随机动力学等学科的论著。

本书可供大气科学、大气物理学专业的研究人员和相关院校师生阅读与参考。

16 开 定价：168.00 元

气象信息系统(现代气象业务丛书)

赵立成 主编

该书内容涵盖了目前气象行业信息技术部门所从事的主要业务工作，包括：气象通信系统、高性能计算系统、气象资料管理与服务系统，以及若干应用系统(如：全国电视会商系统等)。全书由浅入深、由点到面地介绍了这些业务领域的工作范围、内容、结构、流程、相关技术以及标准规范等。通过阅读本书，读者可以较深入地了解气象行业信息技术部门主要业务工作的特点、方法、流程和所涉及的技术范畴。

16 开 定价：50.00 元

天气雷达及其应用(现代气象业务丛书)

李柏 主编

该书深入浅出地介绍了多普勒天气雷达的基本原理及观测方法，详细阐述了多普勒天气雷达在定量估测降水、强对流天气等方面的业务应用，并对多普勒天气雷达应用的发展趋势和发展方向做了介绍。本书共九章，以新一代天气雷达业务应用为基础，首先介绍了新一代天气雷达的基本原理，系统结构、功能、技术指标和观测模式；其次是多普勒天气雷达数据质量控制，多普勒天气雷达产品与算法，典型天气系统雷达回波特征个例的综合分析，其中天气系统的雷达基本图像识别、天气雷达定量估测降水、强对流天气的天气雷达探测和临近预报、典型天气系统雷达回波特征个例的综合分析是本书的重点；最后对新一代天气雷达应用的发展方向做了阐述。每章后都附有参考文献，以便于查阅和进一步探究。

该书适合作为天气雷达、气象业务和科研人员的参考用书，也可作为相关专业人员的培训教材。

16 开 定价：86.00 元

大气涡旋动力学

刘式达 等著

该书力图从涡旋速度场的奇点(即涡旋中无风点)的性质来定性判断这些涡旋的二维、三维结构。书中从力的平衡的角度，求出这些涡旋的速度表达式，利用微分方程定性理论来分析这些涡旋的奇点性质。由此在应用方面，能利用气象上常用的控制参数(水平辐合辐散、涡旋强度等)来判断大气涡旋的结构，从几何和拓扑上将涡旋加以分类。本书在学术上是将涡旋的机理和奇点(无风点)的性质相结合。书中研究结果表明，利用可测的水平辐合、辐散、涡度等物理量，能了解天气分析中的涡旋的三维整体结构及其性质，可对天气预报提供足够的信息和科学依据，有利于提高对大气涡旋在天气预报中的作用及形成机理的认识。

该书是国内第一本论述大气涡旋的专著，也是第一本用定性分析的方法论述大气涡旋的著作，可供从事大气、海洋、天气预报的科技工作者和有关院校师生参考。

16 开 定价：48.00 元

气象出版社网址：http：//www.cmp.cma.gov.cn，E-mail：qxcbs@cma.gov.cn
发行部电话：(010)68406961/9198/9199/8042，传真：62175925