多维、精细粒度高性能计算资源管理系统的设计和在气象业务中的应用实践

顾文静 沈瑜 李娟 孙婧

一种应用于中国气象局高性能计算机系统上的多维、精细化资源管理系统，不仅可以从CPU使用、计算和存储资源的使用、模式、作业等常规角度统计了解系统的运行情况，同时提供对特定队列、特定业务/科研用户或用户群的使用分析，提供对系统全天24小时时段运行情况的使用分析，提供作业等待情况的基本分析等功能。

高性能计算系统是气象业务及科研应用的重要的基础设施平台，该系统对资源统计的大量数据进行多维度、多样化统计分析，构建精准、完善的数据服务，为用户提供高效、标准、可信的应用支撑平台，在保障业务稳定运行的同时，促进系统资源的效能和效率最大限度地发挥。基于该系统，中国气象局（CMA）建立了气象高性能资源计费机制和分配管理模式，使高性能计算机在CMA的应用管理日趋规范化。

1 系统设计与实现

资源管理系统主体功能如图1所示，开发环境：Eclipse Luna Release（4.4.0）；后台数据库：Oracle 11g；Web服务器：Tomcat8.0。资源管理系统包含6大模块，25个子模块，其中首页和作业运行情况模块，实时展示作业运行和资源使用情况，系统管理员可据此调整系统资源分配调度策略，用户可选择空闲的时段和队列提交作业，从而更合理高效的利用系统资源；用户资源统计模块，便于用户查询本人资源申请、分配、使用和缴费情况，为资源计费缴费提供依据。CPU使用情况，资源综合统计和模式应用模式使用情况模块，实现高性能资源多维度、精细粒度统计查询。

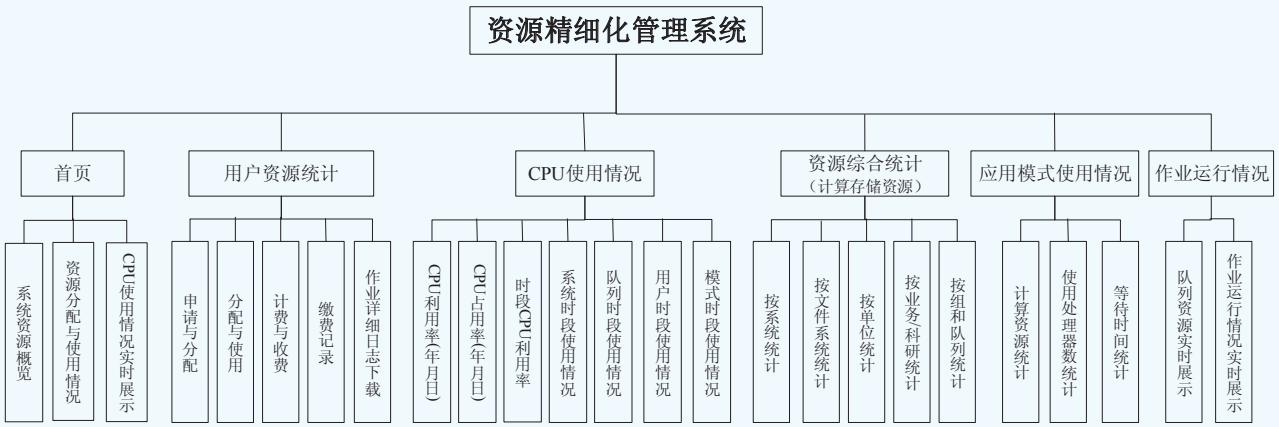

图1 系统模块结构图

1.1 整体架构

资源管理系统基于B/S架构设计多层框架模型，“四个层次”分别为基础设施层、数据层、应用层和表现层，从而保证基础环境的稳定可靠、数据资源的集中管理、应用系统的快速开发、应用系统的集成统一。

1）基础层

基础层包含基础环境运行支撑层和基础软件层。通过对硬件设备、基础系统软件的集成，为本系统提供统一、稳定的运行环境。基础运行环境包括存储系统、高性能计算机、虚拟化云平台以及JDK、Python、shell、Perl和Oracle等组件。

2）数据层

通过对资源数据的统一规划，实现数据的集中存储、数据共享，数据层主要包含源数据和预处理数

据。高性能计算机本地资源管理模块通过基于命令行的功能脚本从高性能计算机系统获取系统运行及用户作业资源使用信息。数据库用于管理资源统计和预处理数据，记账日志经数据解析录入到资源管理数据库。

3）应用层

包含系统的四大模块，其中资源信息采集处理模块包括系统资源信息采集、用户作业信息采集。监视数据管理模块包括资源数据解析入库及预处理，资源信息存储管理。资源统计查询模块包括CPU使用情况、作业运行情况、计算和存储资源统计查询。配置管理模块包括用户权限配置和采集任务配置。

4）表现层

系统主要通过Web页面以及相关的桌面程序为用户提供服务。优良合理的页面布局和人性化的操作方式，清晰地展现用户所关心的信息。

1.2 数据处理

IBM Flex P460高性能计算机系统总理论峰值达1759TFlops；源数据表（job表）总记录近4000万条；日增量约9万条。

伴随数据量的迅速增长，对于用户的统计请求，如果实时的从海量原始数据中进行统计计算，给系统造成较大压力，并且让用户等待较长时间。从用户角度出发，充分利用Oracle数据库索引技术，提高SQL语句的执行效率，增加检索条件避免全表扫描。同时对海量数据进行统计预处理，形成符合需求的中间统计结果进行存储，对于统计访问请求，能够直接在中间统计结果中进行查询，大幅减少等待时间。

资源管理系统数据库创建13张数据表，其中基础信息表包括用户信息表（userinfo）、部门信息表（organization）、计算机系统信息表（machineinfo）、文件系统信息表（gpfsinfo）、资源分配信息表（allocation）；记账信息表包括作业信息表（job）、存储资源信息表（disk）、CPU使用情况信息表（cpu_usage），实时概览信息表（realtimeinfo）；对job表和disk表进行统计预处理，生成队列资源信息表（queueu_amt）、模式资源信息表（model_amt）、系统记账信息表（machine_chg）和用户记账信息表（user_chg）；其中记账信息表的可统计属性均与基础信息表建立主外键关系。

1.3 软件框架

1）模型实现

系统采用B/S架构进行设计，基于J2EE的SSH（Struts2+Spring+Hibernate）架构体系进行业务和数据资源的整合及集成，采用多层体系架构，实现设计单元的高内聚、低耦合；采用组件化与插件机制结合进行设计和开发，保证系统强大且灵活的可扩展性、可维护性以及可集成性。

前台页面通过Ajax 向后台的Action发出Post 请求，后台Action 返回Json 格式的数据给请求页面，请求页面利用JQuery 处理返回的Json 格式的数据，并呈现给用户。通过这种方式高效的实现页面对后台数据的无刷新访问，并且将开发过程中的表现层和数据处理层很好的分离。

视图层采用Struts2框架与jQuery框架实现，Struts2提供对Ajax的支持，与jQuery框架配合，可将程序的触发事件直接封装在JS代码中；Struts2根据用户请求调用相应Action控制器，Action调用Service实现业务逻辑处理，将Json对象和参数传入到了后台Service层，在Service层封装解析数据对象；数据持久层由Spring和Hibernate整合实现，Service调用Dao实现对数据库的操作，Dao通过调用Hibernate API对持久化对象进行操作，Hibernate将Dao连接oracle数据库，将配置文件装载到applicationContext.xml中，最终实现对数据库的增删改查操作。

2）图表生成

本系统使用JQuery+HighCharts来实现资源统计数据图表绘制，并提供下载功能。资源统计数据是时间序列化的数据，通过动态曲线能够更直观地显示数据在一个时间段中的变化过程。

HighCharts是一个非常流行、界面美观的纯Javascript 图表库，能够很便捷地在Web 网站或是Web 应用程序中提供直观、交互性的图表，并且免费和开源。

2 资源精细化统计分析

资源管理系统下设25个子模块，对CPU使用情况、计算和存储资源、队列及作业情况进行了多方面精细化统计查询，下面就其中几个典型模块的应用情况做详细分析。

2.1 首页

首页作为资源管理系统的概览，对所有用户开放访问权限，其中包括CPU使用情况和作业运行情况的实时展示、各单位资源分配与使用情况和模式资源使用情况展示。

CPU使用情况和作业运行情况的实时展示方便系统管理员和用户动态掌握各系统资源使用情况。

2.2 用户资源分配与使用

资源分配与使用情况展示中国气象局各单位和各区域中心计算及存储资源分配和使用情况。根据使

用量与分配量比例判断资源分配的合理性，对各中心用户需求进行充分评估，为下一阶段资源调整提供依据。其包括用户申请与分配、分配与使用、计费与收费、缴费记录和详细日志查询，前4个子模块对所有用户开放，并设置系统管理员、中心管理员和个人用户等权限，系统管理员可查询所有用户信息，中心管理员只能查询本中心用户信息，个人用户可以看到本人名下所有的账户信息，多种权限的设置在保障用户信息安全的前提下为不同需求用户提供了便捷。本模块为中国气象局资源计费机制提供有利依据。

2.3 CPU使用情况

包括中国气象局和区域中心九套系统CPU使用情况和多维度时段CPU核使用情况展示。CPU使用情况包括逐年、逐月、逐日数据的统计查询。逐年、逐月、逐日资源数据统计反映系统在不同时段CPU使用情况，为分析资源使用规律提供有利依据。

时段CPU核使用情况包括系统、队列、用户及业务模式时段使用情况统计。取样间隔精确到秒，通过密集柱状图形象展示了各个时段选定的系统、队列、用户及业务模式CPU核使用情况，统计图（图2）中可以清晰反映那些时间段是作业运行高峰，那些时间段是低谷，为各业务模式系统运行时间调整提供依据，避免资源抢占；科研用户可掌握业务运行间隙提交作业，减少等待时间，提升工作效率。

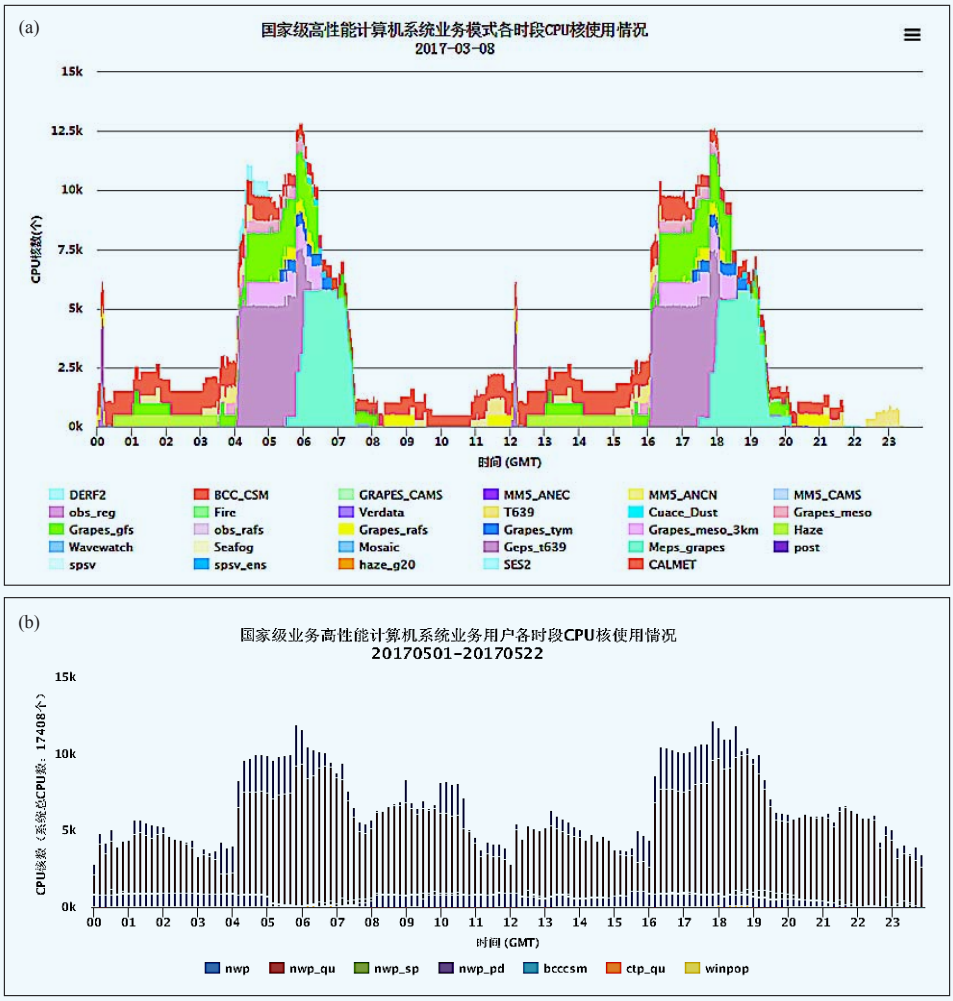

图2（a）业务模式时段CPU核使用情况；（b）业务用户时段CPU核使用情况

2.4 存储资源使用情况

包括文件系统（gpfs）、高性能计算机系统存储资源使用情况统计和用户在各文件系统资源使用情况统计。其中，文件系统（gpfs）和高性能计算机系统使用情况均设计总量和使用量对比展示，清晰反映存储资源使用率；用户统计模块可选择时间查询用户在不同文件系统资源使用情况，为用户调整数据存储策略，合理利用空间提供依据。

2.5 计算资源使用情况

包括单位、用户性质、高性能计算机系统和组及队列等多种统计查询模块。使系统管理员详细了解计算资源使用分布情况，便于调整计算资源分配调度策略，减少资源抢占和资源浪费。

2.6 应用模式使用情况

包含计算资源统计、使用处理器核数和等待时间统计。通过使用处理器核数统计了解各种模式占用计算资源及各种规模作业的比例，据此对模式使用队列、CPU核数、作业数等做一定限制，保证所有模式均可正常运行；等待时间模块设置了30 min—1 h作业和大约1 h 作业详细信息查询，根据作业等待时间及占用资源判断等待是否合理，据此调整资源调度策略，近一步保障业务/科研的稳定运行。

2.7 作业运行情况

包含CPU核使用及作业运行情况和作业运行状态实时展示。CPU核使用及作业运行情况以双轴柱状图对比展示，不仅较单纯数字展示更直观鲜明，同时，通过二者对比可以判断作业运行数和空闲CPU核数是否合理，间接监视系统运行情况，保障系统稳定；运行日志实时提取所有作业关键信息（用户名，作业状态，等待原因）以滚动形式展示在web页面上，避免用户通过命令获取大量复杂信息后的进一步分析提取。

依据资源管理系统，月度、季度及年度定期对CMA高性能计算机系统的进行资源使用和运行的详细分析并编写相应的报告，相应的分析结果为决策者提供决策支持服务。

3 资源的分配与计费

3.1 资源分配

近年来，全国多个地方气象部门在当地政府财政拨款和有关项目支持下，购置高性能计算机系统。自2013年起，依托“气候变化应对决策支撑系统工程”项目，中国气象局陆续在北京、广东、上海、辽宁、湖北、四川以及甘肃和新疆等多个区域中心建设十套高性能计算机系统。

地方气象局部门的高性能计算能力虽然有较大提升，但与仍存在以下一些问题：设备利用率不高、计算能力不能满足当地需求，缺少必要的高性能计算、气象数值模式专业人才。为此，中国气象局每年制定《全国资源分配方案》，依据“本地优先、异地调配”的原则，将一些国家级科研用户分配到省局较空闲的系统上运行，将区域中心一些资源需求较大，本地资源无法满足的业务/科研用户调配至中国气象局系统运行，并充分发挥国家级气象单位在高性能和数值天气预报方面的指导作用，通过资源的预分配，系统管理方将可以主动的控制管理每个用户对不同系统的资源使用。

3.2 资源计费

高性能计算机系统资源有限，为了保障业务/科研稳定运行，最大限度的提高资源的使用效益，中国气象局对系统资源总体遵循“统一计费，分类收取，多用多收，少用少收”的原则。资源单位及计费成本的设定，使得用户可以根据不同情况选择最合适的系统运行，有益于发挥不同系统的效益。服务收费包括计算资源使用费和存储资源使用费，计算资源采用“CPU核小时”作为计量单位，存储资源采用“GB月”作为计量单位。

用户资源统计模块提供用户计算和存储资源不同时间维度下分配与使用、计费与收费和缴费记录查询功能，并对科研用户开放详单查询，提高资源计费透明度，用户可查询本人名下所有高性能计算机账户的资源使用量、应缴金额和已缴金额。该模块作为资源预分配辅助模块，为资源分配的统计提供参考数据，同时实时更新用户分配与使用量，使用户了解本人账户下剩余资源量，促进资源合理使用。

致谢：本文由公益性行业（气象）科研专项（GYHY201106022和GYHY201306062）资助。

深入阅读

陈云芳, 2008. 精通Struts2 基于MVC的Java Web应用开发实战. 北京: 人民邮电出版社.
沈文海, 2012. 网格计算在气象高性能计算领域的应用前景探讨.气象科技进展, 2(1): 48-51.
沈瑜, 李娟, 常飚, 等, 2014．高性能计算机统一资源管理系统的设计与实现．计算机技术与自动化, 33(1): 148-150.
施俊, 2014．基于Struts2+jQuery+JSON 实现Ajax数据显示．电脑知识与技术, 26(10): 6090-6092.
孙婧, 沈瑜, 2015. 气象应用的高性能计算机性能需求推算方法. 计算机技术与发展, 25(6): 206-210.
孙周军, 肖文名, 宋远清, 等, 2008．气象信息实时监视系统改进设计与实现．成都信息工程学院学报, 19(4): 507-511.
王彬, 宗翔, 魏敏, 2008. 一个精细粒度实时计算资源管理系统. 应用气象学报, 19(4): 507-511.

（作者单位：国家气象信息中心）