一个精细粒度实时计算资源管理系统*

王 彬 宗 翔 魏 敏
(国家气象信息中心计算机室,北京 100081)

摘 要

由于相应业务系统软件的缺乏,国家级气象高性能计算机的资源管理措施落后于能力建设的发展。对此,该文提出了一个精细粒度实时计算资源管理系统。系统设计紧密围绕着目前竞争最为激烈的计算资源,采用资源虚拟单元 GCU 作为资源使用的计量单位,屏蔽了不同高性能计算机系统的体系结构差异,实现了计算资源细粒度的统一量化统计。系统可分为用户接口层、资源管理层、HPC 系统层等 3 个层次,根据与网格平台软件不同结合方式以两种方式运行。在国家气象信息中心完成了系统的研发、部署和试验运行,根据试验运行的部分数据进行了用户单位和用户个人的计算资源使用的统计分析。目前,计算资源管理系统成果已成功应用到国家级气象高性能计算机计算资源的业务管理工作中。

关键词: 国家级气象高性能计算机资源;资源管理;GCU;实时;精细粒度

引 言

经过多年建设,国家级气象高性能计算机系统能力建设取得了长足的进展。但相对于运算能力的提高,计算资源管理却由于缺乏先进的软件工具日趋落后,已不能够适应新形势下业务管理的需要[1]。

国家级气象高性能计算机系统的组成日趋复杂,从单个孤立同构计算机集群转变为多个集群系统构成的异构网格环境。服务用户的数量和范围大大扩展,从国家气象中心扩大到所有国家级气象业务和科研单位。

随着天气、气候模式预报精度的提高、物理过程的复杂化,以及集合预报样本的增加,导致模式计算规模越来越大,要求积分时间越来越长,对系统的计算能力和存储资源提出了挑战,给系统资源管理工作带来了一定压力。一方面是虽不断增长但还是有限的资源,另一方面是近于无限的需求,解决资源紧张与用户需求日益增加的矛盾突出。各用户单位均希望提高资源管理水平,满足业务科研工作的资源需要。为此,系统管理者必须要对所掌握的资源使用情况做到心中有数,分配资源、调配资源时才能有理有据,使资源的分配公平、合理,达到最优的使用效果和最高的使用价值。

资源管理具体体现在资源使用记账统计与分配控制两个方面。从目前的资源使用记账上来看,未能实现全局的、综合的、动态的、精细粒度的资源使用记账。现有计算机系统的资源管理没有达到实时记账,粒度粗糙,各自孤立进行,无法获取用户个人/单位在计算系统整体全局的资源统计。另一方面,资源分配还处于较为粗放的阶段,用户开通系统账户之后缺乏足够的约束管理机制,出现了资源分配不公平、不均匀现象:某些用户长时间占用资源,而使有些用户由于得不到足够的资源而无法完成工作;某些时候系统的整体负载非常轻,用户提交作业不多,在系统负载重的时候,用户又集中提交作业,排队等待时间增长,缺乏有力的手段加以约束和引导。这在客观上造成了目前系统的吞吐率不高和资源调度机制不完善等问题。

1 系统设计

1.1 设计目标

基于对国家级气象高性能计算机资源管理现状

* 科技部基础条件平台计划"国家气象网络计算应用系统建设"项目(2005DKA64005)和中国气象局气象新技术推广项目(CMATG2008M07)共同资助。
2007-07-26 收到,2008-04-07 收到再改稿。

和问题的分析，本文提出一个高性能计算资源管理业务系统的设计方案，拟达到 3 个目标。① 实时动态：能够实时动态地跟踪、反映用户对高性能计算机资源的使用情况，并能及时实施资源使用控制策略。② 精细粒度：以一种统一的量化手段描述资源的数量，精确地记录和控制用户资源使用量，从而实现最细粒度的资源记账和分配控制。③ 跨集群（网格）：资源管理范围不仅限于单个高性能计算机系统，而是将所有国家级气象高性能计算机系统都纳入进来，作为一个整体管理、使用全局统一的策略管理。

1.2 设计方案

1.2.1 虚拟计算单元 GCU

高性能计算机资源是指高性能计算机系统为用户提供的能力和服务，包含了多种形式的具体资源，如用户提交作业消耗的 CPU 计算机时、占用的磁盘、内存，调度优先级等。

本文设计的资源管理系统的管理对象限定为目前最主要、也是使用竞争最激烈的 CPU 机时。对于各种异构高性能计算机系统的 CPU 计算资源进行管理，一个核心问题就是如何将其进行适当的抽象，屏蔽厂商、架构、型号、主频等的差异，然后以统一的形式实现精细粒度的量化。

为此，系统设计引入了资源虚拟计算单元 GCU（General Computing Unit），1 个 GCU 相当于目前 IBM 高性能计算机系统 1 个 CPU 小时的计算能力。按照不同系统 CPU 的主频和计算能力，换算出现有运行的计算机系统的 1 个 CPU 小时等于多少个 GCU。比如，神威 32I 计算机系统 1 个 CPU 小时就等价于 0.3 GCU。

通过设计统一的计算单元 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。同时，经过对高性能计算机系统引进购置费用和运行维护费用统计起来，再按照通用的高性能计算机系统的生命期（通常为 5 年），即可折算出 1 个 GCU 在各个系统的成本价格（例如，IBM 高性能计算机系统上，1 个 GCU 可折合为 1.85 元人民币），为资源记账和分配管理提供了分析基础。

1.2.2 目标用户

高性能计算资源管理系统主要为 4 种类型用户提供服务。① 资源用户：主要是高性能计算机系统上的用户，通过提交作业使用计算资源。② 资源使用者组织负责人：资源使用者所属组织领导或项目（课题）负责人。③ 资源系统管理员：高性能计算机系统的管理人员，建立适当的账户、用户组织结构，基于一定策略实施资源分配，监视用户的资源使用情况。④ 决策者：高性能计算资源管理（领导）者，对资源使用的整体宏观情况进行了解，决策资源的分配和引进发展。

1.2.3 功能模块及其关系

如图 1 所示，从整体上看，自上而下，高性能计算资源管理系统可分为用户接口层、资源管理层、HPC 系统层等 3 个层次。

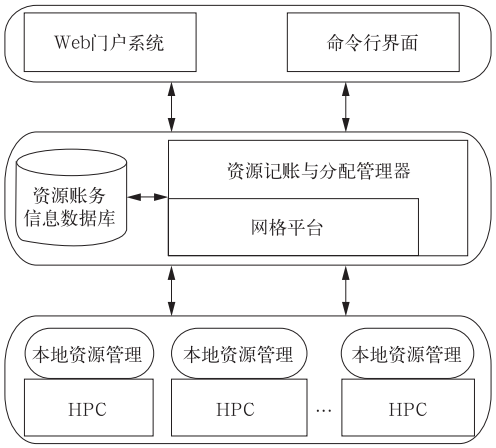

图 1 高性能计算资源管理系统设计方案

Fig. 1 Design scheme of HPC resource management system

用户接口层 用户接口层位于最上层，直接面向用户。考虑不同的用户需要，提供两种接口：Web 门户系统和命令行界面。

资源管理层 资源管理层主要包括资源记账与分配管理器、网格平台、资源账务信息数据库等 3 大部分。其中网格平台充分利用国家级气象高性能计算机管理与应用网络平台建设成果。资源记账与分配管理器是整个系统的核心，包括多个功能模块。

单集群系统 CPU 资源使用记账：实现基于每个用户每个提交作业粒度的动态实时跟踪记录。

资源账户管理和操作：资源账户的创建、删除、修改、锁定/解锁等，资源账户的组/用户管理，GCU 数的预分配和计算等。

网格环境的资源记账：在全局用户一致性管理的基础上实现同一资源账户在不同计算集群使用的统一记账，支持跨集群的资源使用、调度与协商等功能。

资源分配管理：以 GCU 为计算单元对资源用户进行资源分配。

计算能力的分配规划:根据资源使用的历史记录,判断下一季度、下一年度或更长时间的计算能力的供需情况,拟定下一季度、下一年度或更长时间的计算能力分配规划方案。

资源引进决策支持:对于资源能否满足需要,是否有必要引进购置更大能力的计算机系统,提供决策支持和数据依据。

资源账务信息数据库通过关系数据的形式记录了资源账户、资源使用记账、分配管理等信息。

HPC 系统层　最下面的层次是国家级气象部门的各个高性能计算机系统,包括 IBM 高性能计算机系统、神威新世纪集群系统、IBM SP 系统等等。

高性能计算机系统的本地资源管理系统一般是指本地的作业管理器,如 IBM 高性能计算机系统上的 LoadLeveler,神威新世纪集群系统上的 PBS 等。资源管理层主要通过与本地资源管理系统的直接通信,或者是通过网格平台,实现各种资源管理功能。

2　系统实现

2.1　工作基础

资源管理系统的设计和实现充分借鉴了 GOLD 开源技术。GOLD 系统[2-4]是一个开源的资源分配管理器,由美国的太平洋西北国家实验室 PNNL 研发。目前已经在 PNNL 管理主要的集群,作为业务运行的资源分配平台,实现了生产业务使用。此外还在美国十几个计算中心和研究机构进行了应用。

2.2　实现进展

经过半年多的工作,在国家气象信息中心完成了资源管理软件系统的研发、部署和试验运行:① 基于 GOLD 系统,实现了资源记账、分配管理的基本功能,包括集群系统计算作业记账、资源账户管理、用户-组织的管理、分配、查询等;② 资源记账、分配管理的基本功能以命令行的形式提供;③ 数据库选用了开源的 PostgreSQL 数据库技术[5],建立了账户、用户、组织、计算机系统、作业记录、记账、分配等关系表;④ IBM 高性能计算机系统的 3 个分区、神威 32I、神威 32P 和 IBM SP 系统上部署安装了资源管理软件系统,实现了与 LoadLeveler 及 PBS 的集成;⑤ 整理、更新了国家级气象各高性能计算机系统的用户信息,使用了统一的 UID 和 GID,加入 GOLD 数据库中;⑥ 建立了单位(项目)-个人的两层管理机构,即司局级单位账户和个人资源账户;⑦ 2006 年下半年,在 IBM 高性能计算机系统、神威新世纪系列机群系统上,陆续进行了试验实时运行;⑧ 分配策略:按照可用资源 GCU 数总量的 200%,平均分配给各单位(项目);按司局级单位资源账户分配,个人用户只能使用所属司局级单位(项目)账户分配的资源;资源的时效性为一季度,允许超额度使用。

3　测试分析

目前该资源管理系统已经在主要的国家级气象高性能计算机系统上运行,下面根据测试运行期间记录的部分数据(数据采样时间是 2006 年 6—12 月)进行简单的统计和分析**。

3.1　基于用户单位统计的计算资源使用情况

图 2a 给出了主要计算资源用户单位的计算资源使用总计和对比情况。从排名上看,依次为中国气象科学研究院(CAMS)、国家气象中心(NMC)、国家气候中心(NCC)、网格资源共享项目(NMCG)[6-7]、国家卫星气象中心(NSMC)、国家气象信息中心(NMIC)。排名第 1 的中国气象科学研究院在测试运行期间共使用了 1386256.41 GCU 的计算资源,占总量的 64.52%,国家气象中心和国家气候中心分列第 2,3 位,分别使用了总量的 17.86% 和 16.27%,国家气象信息中心只使用了 4036.50 GCU 的计算资源,列最后一位。

图 2b 给出了主要用户单位提交作业的总计情况和对比情况。从排名上看,依次为国家气象中心(NMC)、中国气象科学研究院(CAMS)、网格资源共享项目(NMCG)、国家气候中心(NCC)、国家卫星气象中心(NSMC)、国家气象信息中心(NMIC)。提交作业最多的国家气象中心共提交了 120081 个作业,占总作业数的 64.09%,中国气象科学研究院提交作业数也比较多,有 51673 个,占总作业数的 27.58%,国家气象信息中心最少,只提交了 452 个作业。

3.2　基于用户个人统计的计算资源使用情况

表 1、表 2 分别给出了计算资源使用的个人用户的统计结果。

**本章统计分析使用的数据是系统试验运行期间记录的部分数据,未覆盖全部计算资源的全部运行时间,仅供参考。

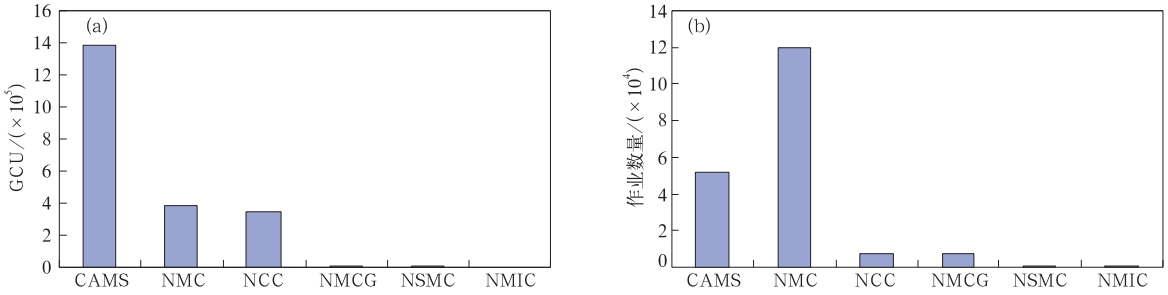

图 2 主要用户单位的计算资源使用(a)及作业提交情况(b)

Fig. 2 Computing resource usage (a) and job submissions (b) by major user organizations

表 1 计算资源使用最多的(前 10 名)用户

Table 1 Top 10 computing resource consumption users

排名	用户单位账户	计算资源使用量(GCU)
1	CAMS	199284.87
2	NCC	136981.96
3	CAMS	126596.98
4	CAMS	103465.45
5	NMC	93645.05
6	CAMS	92602.38
7	CAMS	89232.09
8	CAMS	86046.34
9	NCC	77960.39
10	CAMS	74173.02

表 2 提交作业数最多的(前 5 名)用户

Table 2 Top 5 job submission users

排名	用户单位账户	提交作业总数量
1	CAMS	21044
2	NMC	17061
3	NMC	15926
4	NMC	13207
5	NMC	11711

3.3 基于计算作业统计的计算资源使用情况

表 3、表 4 分别给出了按照作业统计的计算资源使用情况。

表 3 使用计算资源最多的(前 10 名)作业

Table 3 Top 10 computing resource consumption jobs

排名	作业名	用户单位账户	计算资源使用量(GCU)
1	d34n01.107089.0	CAMS	14752.09
2	d35n01.161716.0	CAMS	14752.05
3	d34n01.107090.0	CAMS	14576.37
4	d35n01.161715.0	CAMS	11064.04
5	d35n01.164498.1	NCC	10677.99
6	d35n01.18938.0	CAMS	9152.00
7	d35n01.161718.0	CAMS	9220.04
8	d35n01.159603.0	CAMS	9152.07
9	d35n01.157853.0	CAMS	9152.00
10	d34n01.119438.0	CAMS	9151.40

表 4 各用户单位计算作业使用计算资源情况

Table 4 Computing usage by user organizations

排名	用户单位账户	作业平均计算资源使用量(GCU)	作业运行平均时间/s	使用处理器平均数量
1	NCC	117.03	18053.73	16.01
2	NSMC	60.98	7565.15	20.35
3	CAMS	60.98	5037.99	26.20
4	NMIC	14.38	2670.66	10.37
5	NMC	6.37	400.71	39.00

3.4 数据分析

从以上的资源使用统计数据来看，中国气象科学研究院、国家气象中心、国家气候中心是国家级气象高性能计算机资源最主要的用户单位，占测试统计期间计算资源总使用量的 95%以上。计算资源量使用最多的前 10 位用户分布在中国气象科学研究院(7 个)、国家气候中心(2 个)、国家气象中心(1 个)这 3 个单位。

中国气象科学研究院是测试统计期间计算资源最大的使用者，占总资源使用量的 60%以上，国家气象中心是测试统计期间最活跃的用户单位，提交的作业数量最多，占总量的 63%。

不同单位计算资源使用有着不同的特点。从测试统计数据可以发现，中国气象科学研究院、国家气候中心的平均一次作业持续时间较长、使用资源较多；中国气象科学研究院一次作业平均用时约 84 min，国家气候中心一次作业平均用时约 5.01 h；国家气象中心用户提交作业数量较多，但平均一次运行时间很短，约 7 min/次。这从侧面定量地反映出不同单位资源使用的特点。

科技部网格资源共享项目使用的资源虽然占总量的比重不大，1%左右，但也占据相当的份额，排名第 4，成为国家级气象高性能计算机资源的一个重要用户。随着网格平台的日益成熟和其上越来越多成熟气象应用系统的建成和业务化运行，其所占比

例将不断增加,成为高性能计算机资源服务的重要组成部分。

4　结束语

经过半年多的研发,在国家级气象高性能计算机系统上建立了一个跨异构平台的计算资源管理框架,填补了目前高性能计算机资源管理方面的空白。

试验运行中也发现该资源管理系统在 IBM 高性能计算机系统上运行性能不够理想,急需在记账入库性能、网络配置、数据库服务器配置等方面进行优化。

随着该系统软件开发工作的完成和稳定运行,将制订和实施配套的国家级气象高性能计算机资源管理规定,发挥系统在资源精细化管理、提高资源使用效益等方面的作用。

参 考 文 献

[1] 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. 应用气象学报,2006,17(5):629-634.

[2] GOLD Home Page. http://www.emsl.pnl.gov/docs/mscf/gold/.

[3] Jackson S. Allocation Management Solutions for High Performance Computing. Proceedings of PDPTA 2005, Athens: CSREA Press, 2006: 10-16.

[4] Bodel B, Bradshaw R, DeBenedictus E, et al. Scalable system software: A component-based approach. *Journal of Physics*, 2005, 16: 546-550.

[5] PostgreSQL 8.1.4 Documentation. http://www.postgresql.org/files/documentation/pdf/8.1/postgresql-8.1-A4.pdf, 2006.

[6] 王彬. 国家气象网络计算应用节点门户系统的设计与实现. 气象科技,2006,34(增刊):5-9.

[7] 王彬,魏敏,刘桂英. 基于 NMIC 计算网格平台的 MM5 业务模式共享系统. 2006 年中国气象学会信息技术在气象领域的开发应用研讨会论文集,2006:145-151.

[8] 肖依,任浩,徐志伟,等. 基于资源目录技术的网格系统软件设计与实现. 计算机研究与发展, 2002, 39(8):902-906.

[9] 虞益诚. 基于资源管理的网络技术探究. 计算机应用与软件, 2005, 22(7):69-71.

[10] 郑然,李胜利,金海. 网格资源管理与调度模型的研究. 华中科技大学学报, 2001, 29(12):87-89.

[11] 李春林,卢正鼎,李腊元. 基于 Agent 的计算网格资源管理. 武汉理工大学学报, 2003, 27(1):7-10.

[12] Czajkowski K, Foster I, Karonis N, et al. A Resource Management Architecture for Metacomputing Systems. Proc IPPS/SPDP '98 Workshop on Job Scheduling Strategies for Parallel Processing, 1998: 62-82.

[13] Czajkowski K, Foster I, Kesselman C. Resource Co-Allocation in Computational Grids. Proceedings of the Eighth IEEE International Symposium on High Performance Distributed Computing (HPDC-8), 1999: 219-228.

[14] Foster I. The grid: A new infrastructure for 21st century science. *Physics Today*, 2002, 55(2):42-47.

[15] Foster I, Kesselman C, Tuecke S. The anatomy of the grid: Enabling scalable virtual organizations. *International Journal of Supercomputer Applications*, 2001, 15(3):200-222.

[16] 王涌,肖依,王意洁,等. 元计算系统的一个可扩展层次型资源管理模型. 计算机研究与发展. 2002,39(8):907-912.

A Fine-grained, Real Time HPC Resource Management System

Wang Bin Zong Xiang Wei Min

(*Computer Division, National Meteorological Information Center, Beijing* 100081)

Abstract

In contrast to the rapid development of capability construction, resource management of national meteorological high performance computers is left behind. Absence of operational software in resource management keeps system administrators from having a detailed knowledge of what's going on in national meteorological high performance computers and exerting effective control over resource allocations. Regarding existing problems, a fine grained, real time high performance computer resource management system is proposed. The system is designed to be a real time, fine grained one with cross-cluster (Grid) support. The system works closely with CPU hours resources under keen competition. With the introduction of GCU (General Computing Unit), a resource virtualization unit, to measure computing resources, diversities of computing resources in different high performance computer systems are shielded and fine grained uniform quantitative management is enabled by the system. The target users of the system include resource users, leaders of user organizations, resource system administrators, decision-makers etc. The system comprises three layers, namely, user interfaces, resource management, and high performance computer systems. Resource management layer, the primary layer, can be divided into resource accounting and allocation manager, Grid platform, and resource information database. With open source software from supercomputing centers abroad, Grid project funded by MOST, and RDBMS employed, the system has seen an implementation, deployment and experimental running in National Meteorological Information Center. Fundamental functions of resource accounting and allocation management have been implemented, including cluster system job accounting, resource accounts management, management, allocation and query of user and organizations, providing command line interface for users. PostgreSQL database technology is adopted as the resource information database, on which accounts, users, organizations, computer systems, job records, accounting and allocation relation tables are created. The software system has been deployed into the three partitions of IBM high performance computer system, Sunway 32I cluster, Sunway 32P cluster, IBM SP system, working with LoadLeveler, PBS. Information of users on national meteorological high performance computer systems have been sorted and updated, resulting in uniform UID and GID, and inserted into databases. Two layers of management, organizations (projects) and individuals, are established. Computing resources are evenly allocated to user organizations according to 200 per cent of the total available resource in terms of GCUs. Only resources allocated to their department can be used by individual users. The validity of resources are set to a season. Overdraft is allowed. Based on partial data collected during experimental run, initial statistical analyses are made to probe resource usage by user organizations and individuals. At present, the high performance computer resource system has been put into operational run and successfully applied to operation management.

Key words: national meteorological high performance computer resources; resource management; GCU (General Computing Unit); real time; fine-grained