高性能计算机在华南气象行业的应用研究

张恩红,尹海燕
(广东省气象探测数据中心,广东 广州 510641)

摘 要:为了提高产能,近些年来各行业都在建设高性能计算机系统。该文提到的高性能计算机系统是指中国气象局在华南区域气象中心建设的 IBM 高性能计算机子系统。重点阐述了如何优化计算资源和存储资源的配置与作业调度管理系统,使得系统在华南气象数值预报模式计算中提供较高的运行效率、较高的节点使用率。为了解决海量数据的传输与数据处理,及时准确输出数值预报产品,采用对用户类型的划分、计算节点分组的设计、存储资源独享与共享结合的方法。从系统运行结果来看,取得了显著的成效,提高了存储空间的使用效率,提高了节点的使用率,改善了网络的使用环境。不同类型用户的计算节点可用率提升 157% 至 274%;存储的总需求量比旧系统少了 100 T,提高了 40% 的使用率,数据传输总量减少 55 T,只占旧方案的 45%。

关键词:高性能计算机;气象应用;loadlevel;共享存储;海量数据
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2020)12-0187-05
doi:10.3969/j.issn.1673-629X.2020.12.033

Research on Application of High Performance Computer in Meteorological Industry in South China

ZHANG En-hong, YIN Hai-yan
(Guangdong Meteorological Observation Data Center, Guangzhou 510641, China)

Abstract: In order to increase production capacity, various industries have been building high-performance computer systems in recent years. The high performance computer system mentioned refers to the IBM high performance computer subsystem built by the China Meteorological Administration in the South China Regional Meteorological Center. We focus on how to optimize the allocation of computing resources and storage resources and the job scheduling and management system, so that the system can provide higher operation efficiency and higher node utilization rate in the calculation of meteorological numerical forecast model in South China. In order to solve the problem of mass data transmission and data processing and output numerical prediction products timely and accurately, the methods of dividing user types, grouping computing nodes, and combining exclusive and shared storage resources are adopted. From the running results of the system, remarkable results have been achieved, the use efficiency of storage space has been improved, the utilization rate of nodes has been improved, and the use environment of the network has been improved. The availability of computing nodes for different types of users has increased by 157% to 274%. The total demand for storage is 100T less than that of the old system, an increase of 40% of the utilization rate, and the total amount of data transmission is reduced by 55T, accounting for only 45% of the old scheme.

Key words: high-performance computers; meteorological applications; loadlevel; shared storage; massive data

0 引 言

高性能计算机自从 20 世纪 70 年代问世以来,国内外的气象行业都是高性能计算机应用大户,气象业务的需求也促进了高性能计算机的迅速发展[1-5]。随着高性能技术的发展,气象行业建设的高性能计算机系统性能也逐步发展,从 20 世纪 90 年代的银河 II、神威 I 到神威 4000、IBMP460,再到曙光系列[6-10],计算能力呈指数级增加,为气象行业的数值预报发展提供充分的保障。除此之外,还使用了大量的社会上高性能计算资源,如天河 I、天河 II 等。气象行业数值预报的发展离不开高性能计算机,20 世纪初挪威科学家 Bjecknes[1] 教授提出数值天气预报理论思想,直到二次世界大战出现了大型计算机后,才真正成功地制作出了世界第一张成功的数值天气预报图,花了几十年

收稿日期:2020-02-15 **修回日期**:2020-06-18
基金项目:国家自然科学基金(41805096);广东省科技计划项目(2018B020207012)
作者简介:张恩红(1977-),男,硕士,高级工程师,从事高性能计算机管理与应用、海量数据归档存储管理及数据服务和数据加工处理等的研究。

的时间,最终还是依赖于计算机[11]。自此以后,随着高性能计算机的发展,气象数值预报也得到飞速的发展,全世界的气象行业都在建设高性能计算机系统,NCAR、ECWMF、CMA、MET office 等都建设有超过 5000 TFLOPS 的计算能力的高性能计算机系统,为各国数据预报业务的计算提供了大量的计算资源。随着华南地区社会经济的发展、一带一路的规划以及粤港澳大湾区的建设,华南区域气象中心需要提供大量的数值预报产品,对高性能计算机的需求量是巨大的。如何提供这么大数量的计算资源以及相应的作业调度管理是管理人员和技术人员面临的重大挑战。很多学者和技术人员研究了高性能计算机与气象业务应用的结合技术,有的学者研究了高性能计算系统设计的合理性[12-15];有的技术人员分析了高性能计算系统配置管理的高效性[16-18]。该文着重研究如何高效和充分使用计算机资源,以便发挥高性能计算机的最大效能。

1 对高性能计算机需求背景

华南区域中心具有完全自主数值预报产品研发的能力,包括模式的算法设计、功能实现、性能测试、产品加工等全流程业务。随着社会的发展,国家推出一路一带政策、粤港澳大湾区的建设,对数值预报的需求也成倍增长。从最初的华南区域中尺度(18 km)和南海台风模式发展到如今二十几个数值模式的计算,包括华南区域中尺度 3 km、1 km,一带一路模式、粤港澳模式等。对高性能计算能力从几十个节点到几百个节点的发展。纯业务的需求(不包括科研的需求,科研的需要一般是业务的 3 倍以上)不同时段对计算资源的需求如表 1 所示。

表 1 业务账号对计算机节点和模式运行时次的需求

序号	业务名称	运行模式	运行时次	启动时间	所需 CPU 核数
1	南海台风模式 1	GRAPES	00	5:15	1 024
			06	11:15	1 024
			12	18:45	1 024
			18	23:45	1 024
2	南海台风模式 2	GRAPES	00	7:30	1 024
			06	13:00	1 024
			12	22:00	1 024
			18	1:00	1 024
3	华南中尺度模式 1	GRAPES	00	5:02	1 024
			12	17:02	1 024
4	华南中尺度模式 2	GRAPES	00	7:05	1 024
			12	19:05	1 024
5	华南中尺度模式 3	GRAPES	00	5:30	1 024
			12	17:30	1 024
6	短临预报系统	GRAPES	逐小时	整点 15 分	1 024
7	一带一路区域模式	GRAPES	00	6:00	256
			12	18:00	256
8	南海区域模式	GRAPES	00	5:30	256
			12	17:30	256
9	风能模式	GRAPES	12	21:00	1 024
10	CAMX 空气质量数值预报	CAMX	15:00-24:00	15:00	128
			03:00-07:00	3:00	128
11	陆地和海面交通气象的雾预报业务	GRAPES	20:00-06:00	20:00	1 024
12	GRACEs	wrf	00	11:00	512
			12	21:00	512

续表 1

序号	业务名称	运行模式	运行时次	启动时间	所需 CPU 核数
13	GRACEs	cmaq	00	0:30	1 280
			12	12:30	1 280
14	华南沿海海雾预报系统	Grapes	19:30-22:30	19:30	1 024
15	近岸海浪模式预报系统	Swan	11:10-11:50	11:10	128
			23:50-00:30	23:50	128
16	海浪模式预报系统	Wave WatchⅢ	06:50-11:30	6:50	1
			20:50-01:30	20:50	1
17	南海海洋模式	MOM	21:30-01:00	21:30	1
			07:30-10:00	7:30	1
18	风暴潮模式	storm_surge	22:50-00:50	22:50	1
			06:00-08:00	6:00	1
19	海面风预报系统	Grapes	21:50-22:30	21:50	1
			07:10-08:00	7:10	1
20	9 km 海浪模式	WaveWatch Ⅲ	22:30-00:30	22:30	32
			07:00-09:00	7:00	32
21	海面风融合分析系统	wrf,enkf	逐小时	整点 15 分	32

2　系统设计

2.1　基础设计

华南气象区域中心使用的高性能计算机系统是一套 IBM Flex P460 高性能计算机集群子系统,该系统主要由 P460 服务器(Power7 处理器,芯片:8 Cores,3.55 GHz,8 Floating Point/Cycle,227.2GFlops)组成,计算节点数量为 427 个,总理论峰值达到 391.6 TFlops,物理存储容量超过 700 TB,全系统共计有 CPU 核数为 13 664个,内存总量 58 TB。采用集群配置的模式来管理,集群系统采用冗余方式进行设计,充分保证集群的高可用性和可靠性。高性能计算机系统包括计算节点、存储、登录节点、管理节点、管理网络、Infiniband 网络。本系统之外的系统通过万兆光纤提供数据的共享服务。拓扑结构如图 1 所示。

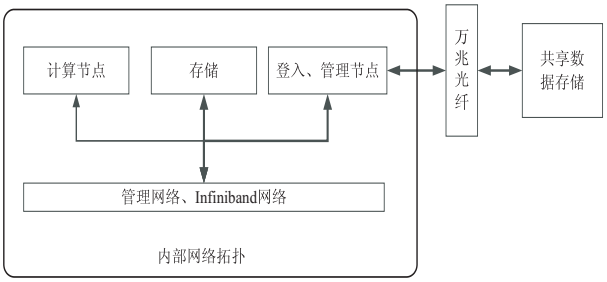

图 1　高性能计算机网络与数据共享拓扑示意图

2.2　作业调度管理设计

为了尽可能满足业务和科研的需求,同时发挥计算资源的最大效能,根据业务的性质和对需求的迫切性,系统将所有用户分为业务账号、重点科研账号和普通科研账号,并且从硬件上也划分出相应的区间,即计算节点按需求进行相应分组,同时,给不同用户组赋予不同的优先级,业务账号高于重点科研账号,重点科研账号高于普通科研账号。业务账号和普通科研账号可以相互共享计算资源,优先使用本区的资源。业务区资源不够用允许抢占普通科研区的资源;业务区的资源闲置,则可以共享给普通科研使用;业务账号的资源在空闲时可以共享给重点科研账号使用,但是重点科研账号的资源不允许其他账号抢占,以确保重点科研资源的最低需求。用户及计算节点资源划分如图 2 所示。

$$\begin{cases} 195 < x < 307; x\text{:业务账号可使用的节点数} \\ 0 < y < 307; y\text{:普通科研账号可使用的节点数} \\ 0 < z < 307; z\text{:重点科研账号可使用的节点数} \\ x + y + z = 419 \end{cases}$$

图 2　高性能计算机账号分类及节点配置

在 loadlevel 的配置部分代码如下所示:

```
# just for special user 195 nodes include 13 largmem nodes
mgroup_1: {
        type = machine_group
        schedd_runs_here = false
        startd_runs_here = true
        MAX_STARTERS = 32
```

```
        #2019/04/12: by zhang enhong
        # Add island and reallocate machines for the group
        machine_list=gza[01-06]n[01-28], gza15n[06-16],gza16n[06-07], gza15n[17-28]
    class=special(32) normal_02(32) normal_01(32)
    }
    #just for normal_01 and special user 112 nodes
    mgroup_21: {
        type=machine_group
        schedd_runs_here=false
        startd_runs_here=true
        MAX_STARTERS=32
        #2019/04/12: by zhang enhong
        # Add island and reallocate machines for the group
        machine_list=gza[07-10]n[01-28]
    class= normal_01(32) special(32)
    }
    #just for normal_02   user 112 nodes
    mgroup_22: {
        type=machine_group
        schedd_runs_here=false
        startd_runs_here=true
        MAX_STARTERS=32
        #2019/04/12: by zhang enhong
        # Add island and reallocate machines for the group
        machine_list=gza[11-14]n[01-28]
    class= normal_02(32)
    }
```

2.3 存储资源共享设计

高性能计算机除了计算资源需要合理调配,存储资源同样需要合理规划和使用。根据业务的特性可知,天气预报中使用的数值预报模式计算都需要大量初始场的数据和观测数据,而且很多模式都需要共同的观测数据和初始场资料,但是通过不同的业务账号运行这些模式。通过需求的调研和业务调整,采用存储独占与共享的模式,即给每一账号分配一个小的存储空间,用户保存私有的数据和本地化的程序,再提供大的存储空间供各个用户共享使用,在此空间中可以存放共同需要的数据,比如:基础观测数据和初始场资料。因此,存储的使用率和数据共享速度都大大提高,也大大降低了对网络的带宽需求。从图 3 的左边模型可以看到每个用户的存储都是独立大存储,从图 3 右边模型可以看到每个用户除了一个独立的小存储,还有个共享巨大存储。

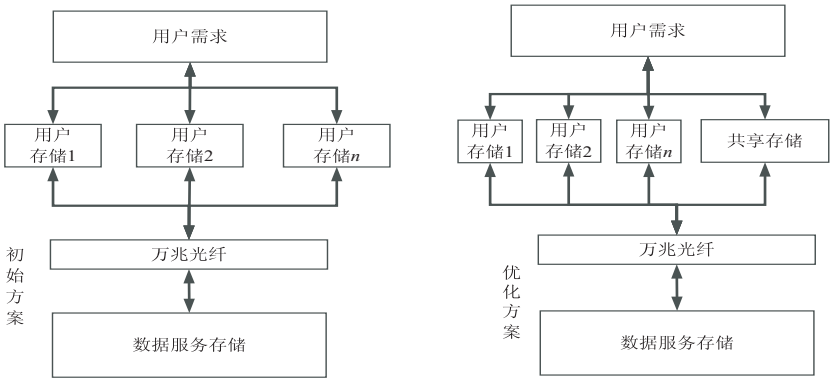

图 3 用户存储分配与共享变化对比

3 业务效果

华南区域中心的高性能计算机的用户主要包括业务用户和科研用户,业务用户 10 个,科研用户 70 多个,其中 40 多个有效用户。日常在线数值预报产品 21 个,日输出数值预报产品 300 多 G,生成十几万个时次的产品;日常科研用户在线作业 20 多个,每日科研产品超过 1 000 G(不提供数据服务,仅作为科研分析使用)。

从图 4 可以看出,业务资源的使用是有阶段性的。对节点的需求量,不同时段对节点需求量是不同的,最少的时次只需要 48 个,最多时次达到 228 个。因此,业务区节点有时候是空闲的,可以共享给科研使用,有时候是不足的,需要从科研区抢占一部分资源,这样既能满足业务的需求,同时也可以给科研用户提供计算节点使用的机会,即科研用户可以在业务闲时提交作业,并且共享业务区的计算节点资源,如 UTC 时间 9-12 时。

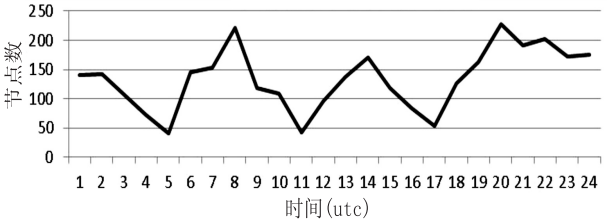

图 4 业务账号计算节点需求不同时次的变化曲线

从表 2 可以看出,存储的总需求量少了 100 T,只需要旧方案的 60%,数据传输总量减少 55 T,只占旧方案的 45%。可见新方案对高性能计算机系统的性能提升是显著的,大大提高了存储的使用率和网络数据传输的效率,同时也缓解了网络带宽的压力。

表 2　新旧方案存储使用对比

方案	用户数	平均存储/T	共享存储/T	平均传输数据量/T	共享传输数据量/T	总存储需求/T	总数据传输量/T
旧方案	50	5	0	2	0	250	100
新方案	50	1	100	0.1	40	150	45

从图 4 可以看出,业务账号在大部分时次需求的计算节点是低于 195 个,有三个时次计算节点是不能满足的。因此,在空闲时段,可以把部分节点共享出来给其他用户使用,在计算资源不足时,可以从普通科研区抢占部分资源以达到业务需求。从表 3 可以看出,业务账号可使用节点达到 307 个,可用率提高到 157%,显然是满足当前的业务需求的;重点科研和普通科研账号业务都可以使用上限节点达到 307 个,可用率提高到 274%,科研账号避开业务繁忙期,可以使用充分展开科研计算。

表 3　新旧方案计算节点使用对比

账号	改造前可使用节点	改造后可使用节点	可用率/%
业务账号	195	307	157
重点科研	112	307	274
普通科研	112	307	274

4　结束语

华南区域中心的高性能计算机系统给华南区域气象中心的数值预报提供充分的计算资源,为华南区域天气预报的计算提供了重要保障。该文简单阐述了华南区域中心的高性能计算的基本情况,重点分析了如何优化高性能计算节点的应用规划和作业调度管理,以便提高计算节点的使用率,提升用户的作业完成的及时性、有效性;如何优化存储资源的分配方法,以便提高存储资源的使用率,减少数据的无效传输,降低网络的负荷。从使用效率来看,当前的方案成效是显著的,不同用户类型的计算节点可用率提高 157% 至 274%;节约了 40% 的存储空间,减少了 55% 的数据传输。

参考文献:

[1] 赵立成,沈文海,肖华东,等. 高性能计算技术在气象领域的应用[J]. 应用气象学报,2016,27(5):550-558.

[2] 孙　婧,沈　瑜. 气象应用的高性能计算机性能需求推算方法[J]. 计算机技术与发展,2015,25(6):206-210.

[3] IBM Blue Gene team. Overview of the IBM blue gene /P project[J]. IBM Journal of Research and Development, 2008,52(1):199-220.

[4] SCHROEDER B,GIBSON G. A large-scale study of failures in high-performance computing systems[J]. IEEE Transactions on Dependable and Secure Computing,2010,7(4):337-350.

[5] FU Haohuan,LIAO Junfeng,YANG Jinzhe,et al. The sunway Taihu Light supercomputer system and applications[J]. Science China Information Sciences,2016,59(7):072001.

[6] 王　彬. 高性能计算技术在气象部门的应用[J]. 计算机工程与设计,2014,35(4):1476-1479.

[7] 王俊超,彭　涛,冯光柳. 曙光高性能计算机在数值预报模式中的应用[J]. 计算机技术与发展,2014,24(10):178-181.

[8] RAMEZANI F,LU Jie,HUSSAIN F K. Task based system load balancing in cloud computing using particle swarm optimization[J]. International Journal of Parallel Programming,2014,42(5):739-754.

[9] RAO Jun,SHEKITA E J,TATA S. Using Paxos to build a scalable,consistent and highly available datastore[J]. Proceedings of the VLDB Endowment,2011,4(4):243-254.

[10] 孙　婧,李　娟,沈　瑜,等. 中国气象局高性能计算机发展历程[J]. 气象科技进展,2018,8(1):50-51.

[11] 张　宇,陈德辉. 谈谈高性能计算机对数值天气预报发展的重要技术支撑作用[J]. 科研信息化技术与应用,2010(4):9-19.

[12] 沈　瑜,孙　婧,李　娟. 中国气象局高性能计算机系统高可靠性设计[J]. 信息安全与技术,2013,4(6):42-45.

[13] 顾文静,常　飚,李　娟. 高性能计算机资源管理系统改进设计与实现[J]. 计算技术与自动化,2017,36(4):104-109.

[14] 张志坚,伍光胜,孙伟忠,等. IBM Flex P460 高性能计算机系统及气象应用[J]. 现代计算机,2016(9):51-55.

[15] 赵春燕,孙　婧,魏　敏. 云及高性能计算集群环境中配置管理系统设计[J]. 计算技术与自动化,2016,35(1):111-116.

[16] 吕　爽,衡志炜,马艳军. 西南区域气象中心 IBM 高性能计算机管理及应用[J]. 高原山地气象研究,2015,35(2):71-76.

[17] 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设与应用[J]. 气象,2011,37(6):764-770.

[18] 许皓皓,李从初,姚浩立,等. 气象高性能计算机故障监控系统的设计与实现[J]. 计算机时代,2017(8):90-93.