DOI:10.13878/j.cnki.jnuist.2016.03.007

朱宏武[1]　尹新怀[1]　罗丹[2]　贺炜[1]　戴泽军[3]

湖南省气象局远程高性能计算环境的设计与实现

摘要

湖南省气象局依托国家超级计算长沙中心,建立了我省第一个远程高性能计算终端用户.针对远程环境的搭建,首先分析了湖南省气象局在高性能计算方面的需求,然后从省级气象部门业务计算需求出发,重点阐述了以高性能计算机集群"天河一号"为计算资源的远程计算环境的系统架构以及主要技术路线与方法.考虑到远程高性能强大的计算能力和省级气象部门作业提交的复杂需求,又进一步设计了湖南省远程高性能任务调度的系统流程以及保障远程计算环境高速稳定的多层面方案.最后也给出了高分辨率中小尺度 WRF 模式在该环境下运行的实际情况,计算速度改善非常明显.

关键词

远程高性能计算环境;系统架构;作业调度系统

中图分类号 TP393

文献标志码 A

收稿日期 2014-06-03

资助项目 公益性行业(气象)科研专项(GYHY201306003)

作者简介

朱宏武,男,博士,主要从事气象信息技术、高性能计算方面的研究工作.
ZhwBUPT@ Gmail.com

1　湖南省气象信息中心,长沙,410118
2　湖南省气象服务中心,长沙,410118
3　湖南省气象科学研究所,长沙,410118

0　引言

随着社会的快速发展,政府、社会和公众对天气预报服务有了越来越高的要求,促使省级气象部门提供更快、更准的预报.虽然当前日益丰富的气象资料为高质量预报奠定了基础,然而为处理这些海量资料,运算区域高时效、高分辨率气象气候数值模式的计算量也在呈几何级巨增,省级气象局有限的计算资源成为影响预报质量的瓶颈.目前国家气象局,北京、广东等经济条件较好的省级气象局已独立购买了高性能计算机解决所需,此类方案具有投资成本过高、运行维护难度较大等缺陷,一般省级部门采用并不现实.值得注意的是,湖南目前已成立国家超级计算长沙中心,搭建了以"天河一号"机群为主的高性能计算服务平台.针对当前形势,湖南省气象局搭建了远程超算高性能计算环境,以此来解决计算资源不足的问题.

1　远程高性能计算环境的平台建设

远程高性能计算环境平台建设是一项面向气象、气候业务的系统工程,相对于本地计算环境的搭建,有以下问题需要解决:

1) 合理搭建融合本地气象业务特色的远程超算环境.与超算中心互联的网络通道不仅有高速的要求,而且需相关的策略保障.如数值天气预报的计算,具有典型高实时性要求,进行区域数值预报时,需要外部提供背景场和侧边界等数据条件,数据量大,特别是在灾害性天气期间,而气象部门资料包含 14 大类、498 个子类、1 500 个细目的气象观探测数据和产品,为满足特定时段对指定资料在传输过程和作业调度的及时处理,需采取从传输设备到调度软件等多种方式的保障,以确保业务执行的轻重缓急.

2) 气象业务通过远程高性能环境进行计算,其作业调度系统的自动化、智能化、模块化程度要求更高.为将远程超算平台融入气象部门业务中,在整合好气象局客户端前台和超算中心后台的同时,还需合理划分本地计算预处理任务和远程后台的并行计算任务,实现气象业务作业调度的灵活控制、调整和监视,该远程调度系统的稳定性和高效性也需要重点考虑.

3) 远程超算环境需要良好的基础设施支撑,其稳定性、安全性、可靠性也有较高要求,包括对各网络节点、通信链路、电源供给等的状态监控,以及故障处理的应急措施,以满足气象局与远程超算中心

的高速互联、气象资料的稳定传输、系统资源的安全访问等需求.

在参考国内外超算中心多处组建方案[1-10]的基础上,搭建了湖南省气象局远程超算环境.本文将在3个方面阐述具体的技术路线,即远程超算环境平台的系统架构、远程作业调度的系统流程和远程超算基础环境的稳定保障.

1.1 远程超算环境平台的系统架构

在参考国内外主流系统架构[2-5,8-10]的基础上,根据气象业务实际情况,设计了本省远程高性能计算环境,主要分为气象数据信息收集网、预警中心局域网、超级计算中心3部分(图1).

1)气象数据信息收集网.主要用来收集省、市、县三级及周边邻省实时气象资料.资料通过高空气象卫星下发,经地面专用网络SDH、MSTP、MPLS、GPRS和公用网络Internet等不同方式由网关设备汇入局域网.

2)预警中心局域网.其主要任务包括接收来自气象源的数据,预处理后导入其存储介质;将远程计算相关数据传输至超算中心,并在计算完成后取回送至SWAN、Micaps等应用系统和Web、FTP及文件共享服务器.

局域网核心层由2台采用了VSS虚拟化技术的万兆交换机组成,各气象终端业务机之间通信带宽保持在千兆以上.省局至超算中心约60 km,为保障高带宽,网络通道介质采用租赁单模裸光纤的方式实现,目前实现端与端间千兆互联.气象数据经省局边界防火墙通过源地址NAT转换,以静态路由的方式实现与远程超算平台的互联.

为了应对资料网络传输中突发性、实时性要求,保障资料处理的轻重缓急,不同资料在特定时期传输时实施了动态的处理策略.目前对该情况处理有多种策略,如先进先出队列(FIFO)、优先级队列(PQ)、定制队列(CQ)、公平队列(FQ)、基于类的加权公平队列/低延迟排队正文(CBWFQ/LLQ),可根据策略的特点,合理处理调度策略.各传输队列策略特点如表1所示.

不同资料在重要的网络设备中设置了相应的服务质量(QoS)策略,特别是重大性、关键性、转折性灾害性天气时就应对传输资料进行合理队列调度,利用访问控制列表(ACL),对数据流进行匹配,如图2所示.

建立CBWFQ/LLQ策略,将气象资料重要程度分为金、银、铜和其他多种类别,然后在网络接口上指定带宽值,配置流量整形,并在接口启用CBWFQ策略,以满足网络传输方面突发性、实时性的要求.

3)超级计算中心.超算中心"天河一号"大型机系统基于集群设计,Linpack值达到772 Tflops,通过InfiniBand架构80 Gbps私有高速网络实现结点间互连,对外提供单一系统映像,能很好地支持大规模、复杂气象业务系统的计算.为满足气象部门大型并行计算的需要,系统提供了OpenMP共享内存和MPI分布式内存2种并行运算方式.为了保障气象计算处理高实时性要求,超算中心为省局安排了固

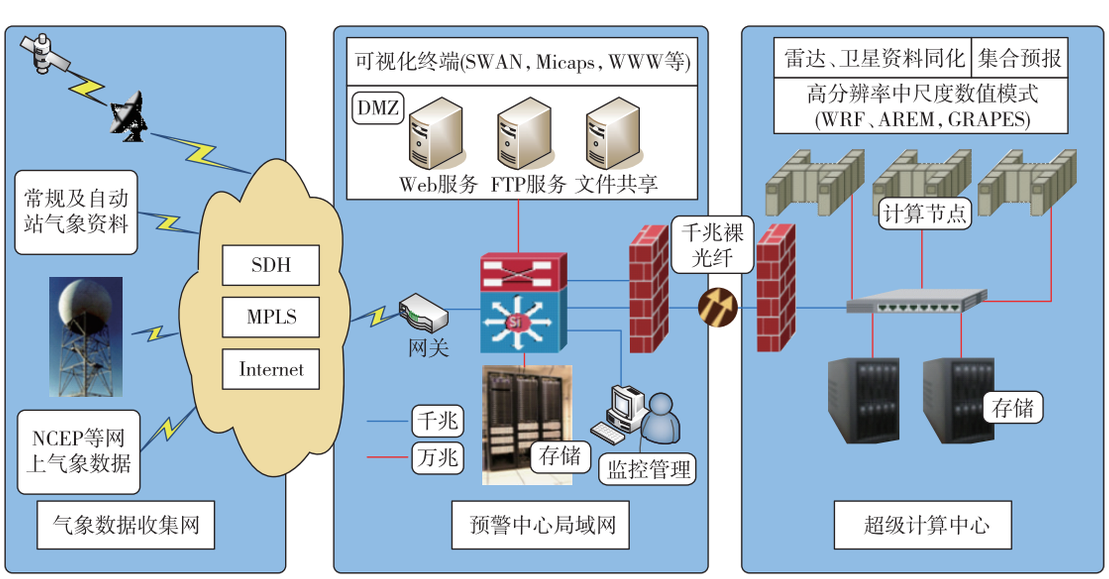

图1 湖南省气象局高速网络通道构成

Fig. 1 High-speed network channel in Hunan provincial meteorological bureau

表 1　气象资料通信传输任务主要策略

Table 1　Advantages and disadvantages of main scheduling policies for meteorological information communication

策略	默认/最大队列数量/个	优点	缺点
FIFO	1/1	易配置,处理简单、延迟小.	处理速度快,无服务质量.
PQ	4/4	绝对保证了优先级高队列调度,延迟、抖动小.	处理速度慢,可能造成低优先级队列得不到调度.
FQ	16/255	实现了公平性,配置简单.	影响重要资料优先处理.
CBWFQ	16/255	照顾某些流的同时保证了其他流的公平性.	影响处理速度,配置复杂.
LLQ	1/1	绝对保证给定带宽下的数据流的延迟、抖动最小.特别适合实时流的应用,同时保证了其他流的公平性.	影响处理速度,对超带宽的数据包直接丢弃.可与 CBWFQ 联合使用.

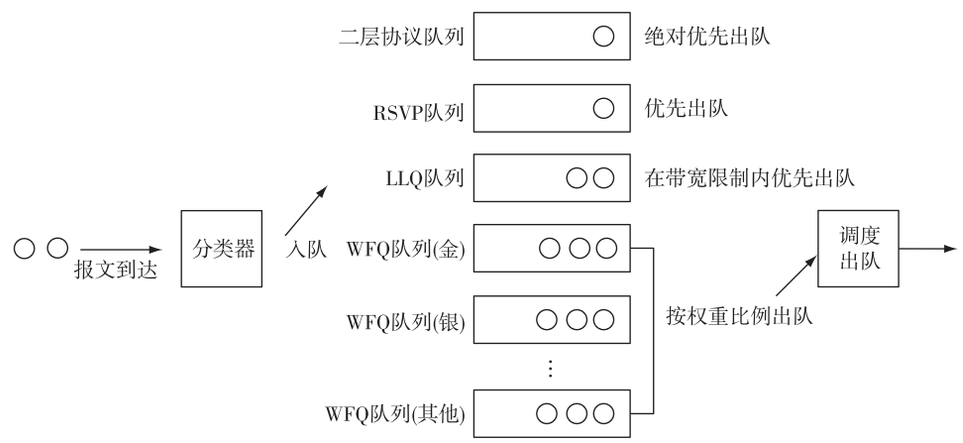

图 2　CBWFQ/LLQ 策略流程

Fig. 2　Scheduling process of CBWFQ/LLQ policy

定的 CPU 可计算资源 36 Tflops,存储资源 30 TB.

1.2　远程作业调度的系统流程

气象业务通过远程高性能环境进行计算,其作业调度系统的自动化、智能化、模块化程度要求较高.为将远程超算平台融入气象部门业务中,在整合好气象局客户端前台和超算中心后台的同时,还需合理划分本地计算预处理任务和远程后台的并行计算任务,实现气象业务作业调度的灵活控制、调整和监视.另外,考虑到作业系统计算基于远程,其调度系统的稳定性和高效性也需要得到保障,此部分保障是设计的重点和难点.根据实际情况,设计了本省远程超算作业调度的系统流程,如图 3 所示.

省级用户通过作业调度系统提交作业至远程高性能计算机运行.作业调度应该经过用户访问、控制,气象、气候模式匹配,调度策略,获取计算、存储资源分配等过程完成作业的运行[4-7].模式及应用所需数据可从本省气象数据环境获取.模式运行所需或产生的大量气象数据可通过模式子系统进行保存和访问.模式运算结果传至指定的共享服务器供业务应用.高效、稳定的作业调度系统是建设的重点,包括如下 4 个子系统:

1) 硬件支撑环境子系统.该子系统包括服务器、操作系统、网络设备、集群管理软件、并行文件系统等,目的是为用户构筑一个虚拟化的、高可用的、高效的支撑环境,该系统有主、备两节点,其主节点是一套基于 Veritas Cluster Server 的计算机集群系统,包含 4 台服务器,作业提交时能够实现 4 台服务器的负载均衡,主节点只要 1 台机器能运行,业务就不会中断.考虑到集群系统故障,设计 1 台服务器进行应急备份,以保障远程作业调度系统的稳定性.集群系统包含有共享存储子系统:集群文件子系统(CFS)、集群全局锁管理子系统(GLM)和集群卷管理器(CVM)子系统,实施了多系统间的协同与负载均衡功能.系统架构如图 4 所示.

CFS 子系统利用 GLM 层全局锁、分布式锁管理器控制集群文件的访问一致性.CVM 协同集群多节点并行访问共享文件卷,协同集群成员关系,保证消息以原子形式传递,保障集群节点通信的低延迟、自动负载均衡、自动检查链路状态等.为提高底层磁盘

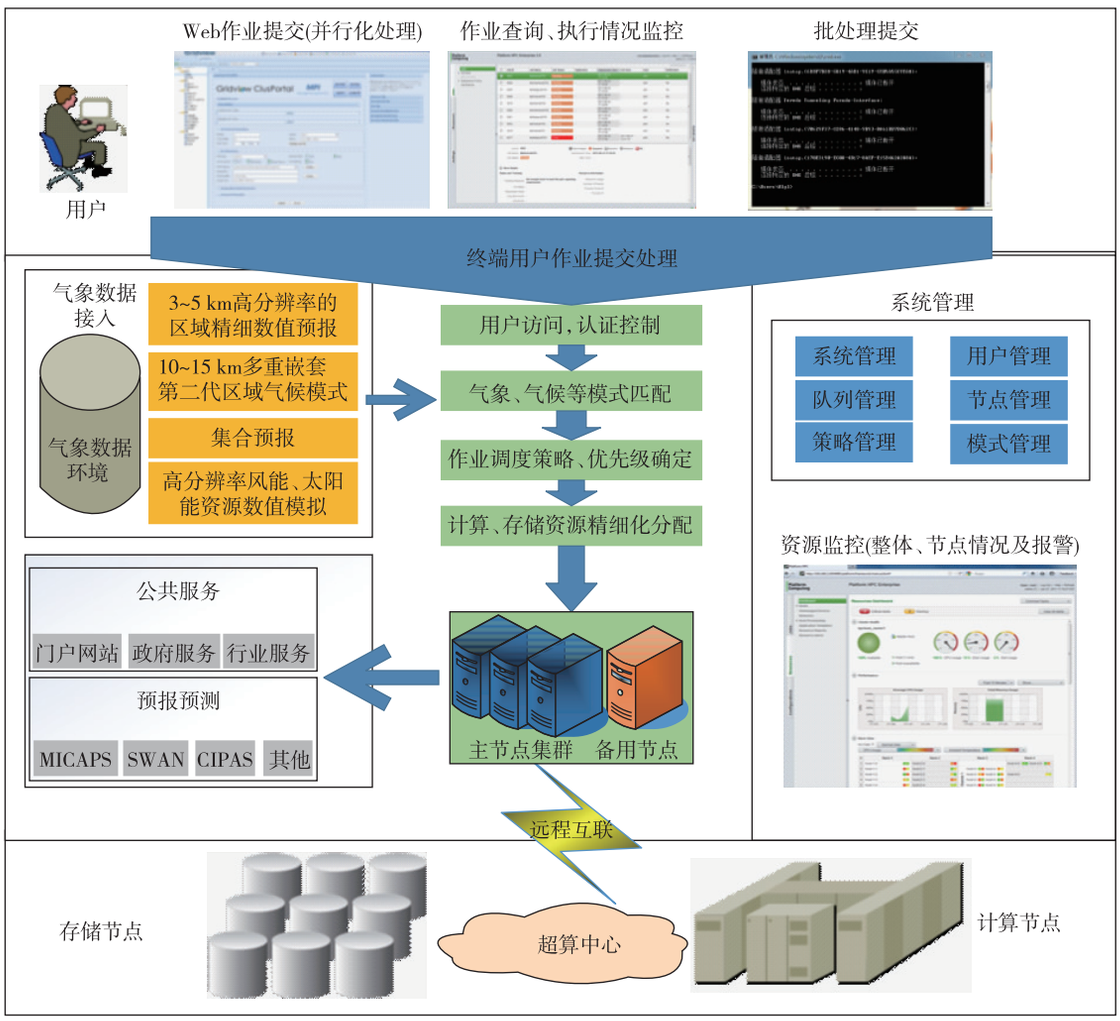

图 3　远程作业系统流程

Fig. 3　System process of the remote job scheduling

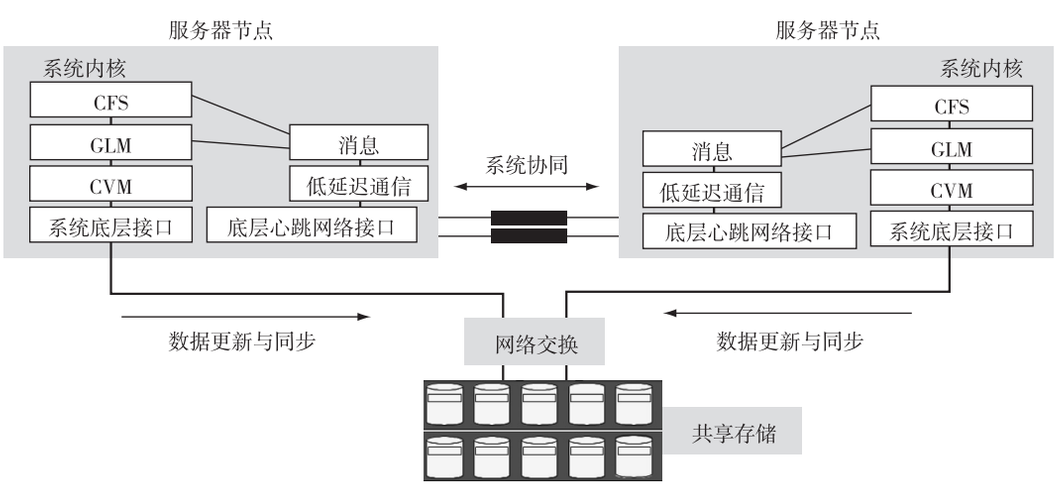

图 4　硬件支撑环境高可用性集群节点系统架构

Fig. 4　System architecture of hardware support environment for HA cluster nodes

利用率,将磁盘分组成多个逻辑卷,并将系统 I/O 分　　配到多个路径上以提升磁盘性能.远程高性能系统

提供 C、Fortran 等多种高级语言以及 MPI 并行消息通信库的高性能并行计算环境,便于用户程序不做修改或做较少修改就能较好地远程运行.

2) 作业管理子系统.该子系统包括作业系统管理、用户管理、队列管理、策略管理、模式管理,以及并行编译器、并行函数库、并行开发环境,还有气象科学计算行业应用软件环境 WRF、GRAPES 等.作业调度系统尽量将复杂的天气、气候数值模式计算提交到超算中心,将用户认证、资源分配以及模式计算所需的背景场、初始场条件等要求不高的计算留在本地.作业调度系统能够设定相应的优先级别,能够根据作业调度系统设置的参数,动态调整运行的程序和分配的资源,保障特定时期处理重要作业程序能够分配到足够资源并优先执行,用户可通过 XMPI 性能分析工具实时查询执行情况,便于程序的优化.

3) 资源监控子系统.该子系统负责对各种类型的资源进行有效调度、管理、监控和维护,具体包括全系统所有功能部件与设备的监测和故障诊断,实现对计算资源和存储资源的使用情况实时记录与统计分析功能,并依据数据分析,从高性能资源整体的角度合理分配、调度资源.该子系统可通过单点登陆实现复杂的多节点/多步骤的作业流引擎的管理,提供资源监控、日志审查功能等.

4) 资料收发及共享子系统.该子系统的任务包括高性能计算所需的各项实时气象资料的收集、处理、传输等,完成作业提交前的预处理工作,另外还需将高性能计算返回的结果及时分发给各业务部门.省级气象部门资料繁多、定时,在灾害性、关键性、转折性天气时对资料处理还分资料处理优越性等级.省级收发与共享系统主要基于自动文件分发系统 AFD 完成.不同于传统文件传输,该子系统底层采用高效的消息机制完成.传输流程如图 5 所示.

该系统包含 2 个主程序:自动消息产生器 amg 和文件分发程序 fd,文件分发是 amg 首先需要寻找源目录中代发的文件,其文件收集主要通过 gf_sftp 检查该文件是否需要预处理,如需预处理则通过 dir_preproc 进程完成,完成后即产生一条消息进消息队列,消息创建成功,会产生应用格式消息目录,格式如下:

files/outgoing/<JID>/<counter>/<creationtime>_<uniquenumber>_<SJC>

其中 JID 表示分发任务的 ID,counter 表示消息计数,creationtime 表示任务创建时间,uniquenumber 表示任务唯一计数号,SJC 表示子任务号.消息产生后,就会将待处理文件拷贝到 fd 程序的目录,fd 主程序仅需观测 amg 消息队列,读取消息的优先级和目的地,通过 dir_check 检查文件格式是否正确,确定无误后将文件通过 sf_ftp,sf_loc,sf_sftp,sf_scp 进程送到目标文件目录,每操作一步都将其日志写入日志消息队列,便于管理和统计.本地或远程气象资料的收发及共享服务采用成熟的 J2EE 平台构建,Linux 服务器系统上的 VSFTP、基于 Tomcat 的 Web 服务及其底层集群文件系统等能够在多台服务器上采用负载均衡的方式完成文件的并行处理,而需要一致性处理的后台 MySQL 数据库采用 HA 方式实现高可用.该系统依照处理资料的类型不同采用了多种负载均衡

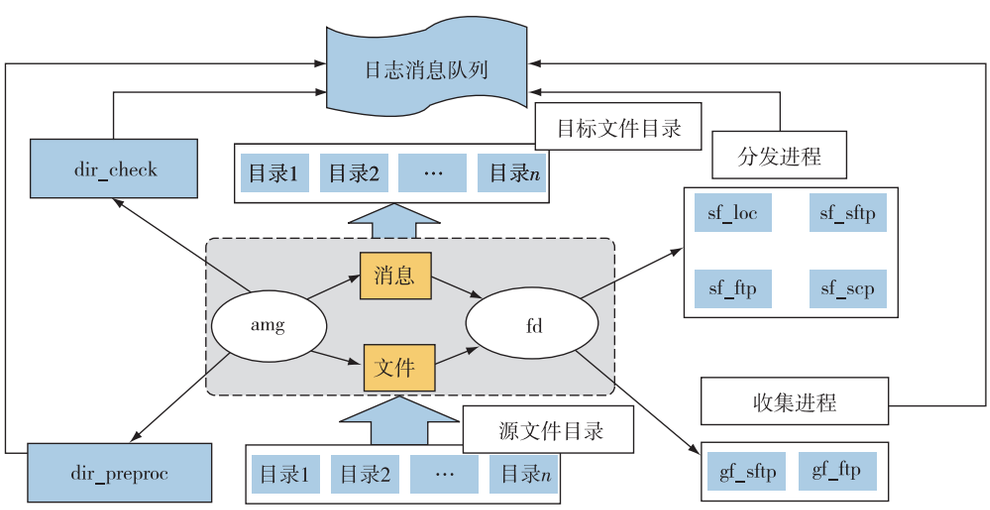

图 5 资料收发及共享子系统处理流程

Fig. 5 Scheduling processes of information receiving,sending and sharing subsystem

方式分配任务,包括轮流均衡(RRS)、加权轮流均衡(WRRS)、最小连结数均衡(LCS)、加权最小连接数均衡(WLCS).通过该种消息方式实现气象资料收集分发,能够高效地对本地或远程文件按优先级不同、分类不同、时间不同进行自动分发,满足了省级气象部门资料传输的要求.

1.3 远程超算基础环境的稳定保障

远程计算环境对基础环境建设稳定性、安全性及可靠性和可管理性的要求也需着重考虑.省局远程环境搭建采用目前较为先进、稳定、安全的解决方案,包括交换机、防火墙、SSL VPN、防毒墙、入侵检测系统(IDS)、网管软件等产品,尽量使各自产品的性能充分发挥,提高整体系统的互通性、安全性、可用性、可靠性、透明性和可管理性.网络安全架构设计如图 6 所示.

为了保障平台通信安全,在部署相关设备时采用了分段设置、分段控制的方式.将安全大体分为如下 4 个区域:

1) 超算用户至省局核心交换机.该段采用基于 TSM 代理提供安全接入控制实现访问终端的安全,解决网络准入问题.

2) 服务器到核心交换机.采用 IDS 策略对网络中数据包主动分析,多重检测实现数据安全,能有效防范对蠕虫的攻击以及隐含在加密数据流中的攻击.

3) 核心交换机到防火墙.通过网络端口安全策略和防毒墙的联动方式实现此段安全.端口安全策略的设置能够有效地从网络源端口数据控制.通过防毒墙能够主动发现网络包中的恶意脚本、病毒,以保障数据包的安全.

4) 防火墙到超算平台防火墙.省局的防火墙到超算平台做专门区域,并设置相应白名单,严格控制网络的访问.

高可靠性和可管理性是气象业务化运行的重要指标,主要通过资源冗余灵活设置网络协议、策略,且配合专业软件实现.核心网络设备采用了双设备、双电源冗余模式.远程通道应急时也可以采用 VPN 方式通信.为保障可靠性,机房也配备了精密空调、多路 UPS 电源等基础设施.

2 远程高性能计算环境的应用

远程高性能计算平台的搭建,极大地增强了气象试验平台计算能力.原来省局计算机集群运算 9 km 的 WRF 模式超过 3 h,更小尺度的计算是不切实际的,而现在基于远程的超算环境,运算时间大幅缩短.现运行高分辨率 4 km 的 WRF 模式仅需40 min,运算结果与实况也很接近(图略).

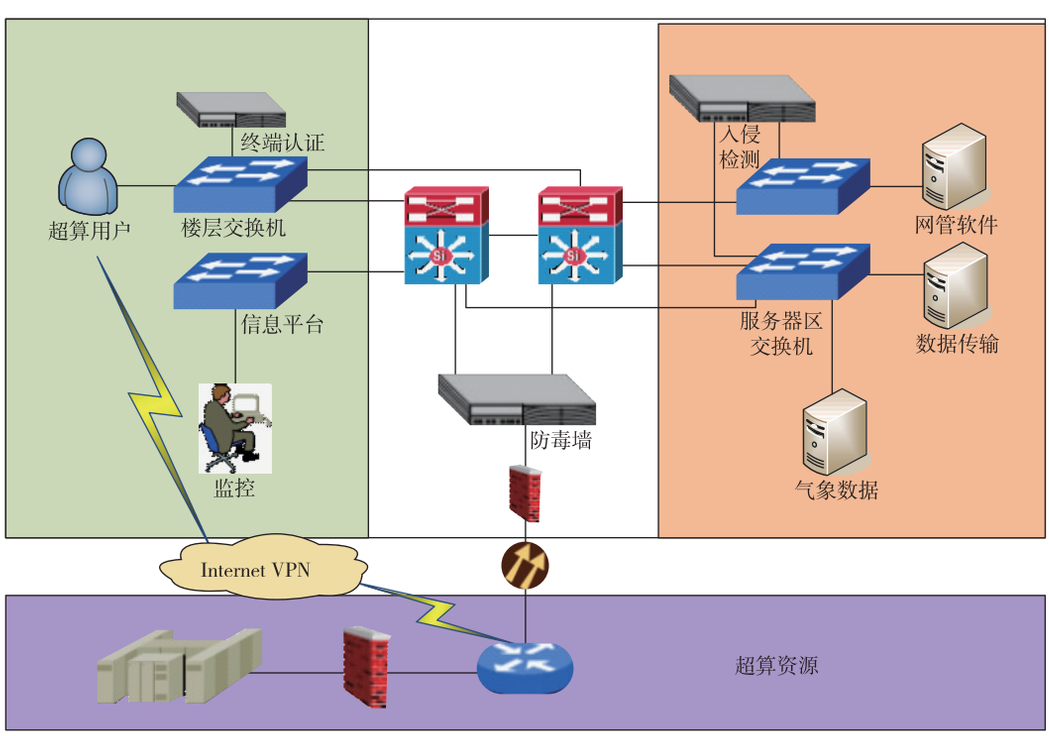

图 6 湖南省气象局网络安全通道设计

Fig. 6 Network security architecture in Hunan provincial meteorological bureau

3 结论

湖南省气象局以提高省局内部高性能计算环境需求为牵引,屏蔽了资源异地性和网络的异构性,已经搭建了一套规范的远程高性能计算应用环境.重点阐述了建立以高性能计算机为计算资源的远程计算环境的系统架构及采用的主要技术路线与方法.考虑到远程高性能强大的计算能力和本省气象部门作业提交的复杂需求,又进一步设计了本省远程高性能任务调度的系统流程.最后也给出了高分辨率中小尺度 WRF 模式在该环境运行的实际情况,计算速度改善非常明显.平台建成具有重要意义,主要包括如下两方面:

1)远程高性能计算环境为省局计算能力提升了一个数量级,使得湖南省气象局计算能力跃居全国前列.应用该环境能够更快地运算高时效和高分辨率的数值模式,为更快更准分析重大性、关键性、转折性天气形势提供了可靠依据,也为即将开始研究的嵌套区域气候模式、集合预报等科研项目夯实了计算的硬环境,并为未来气象海量信息处理分析、重大科技项目攻关等提供了科技支撑.

2)省局远程高性能计算环境的建设既体现了国家重要资源的集约化建设与高效应用的原则,也符合国内外对高性能计算集中建设与应用的发展方向,是一种资源节约、高效利用的理想模式.此次成功尝试,也为未来其他集约化资源利用积累了经验,同时也可为其他省市建立高性能计算环境提供一种思路.

参考文献

References

[1] 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
ZHAO Wei, LI Minghao, TANG Yuanming, et al. Design of Liaoning meteorological network computing application system[J]. Meteorological Monthly, 2009, 35(12): 133-138

[2] 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
ZONG Xiang, WANG Bin. Design and practice of national meteorological HPC management and application network platform[J]. Journal of Applied Meteorological Science, 2006,17(5):629-634

[3] Malawski M, Gubała T, Bubak M. Component-based approach for programming and running scientific applications on grids and clouds[J]. International Journal of High Performance Computing Applications, 2012, 26(3):275-295

[4] 刘俊铖,黄瑞芳,杨波,等.基于工作流的气象高性能计算用户环境[J].计算机工程,2010,36(8):278-280
LIU Juncheng, HUANG Ruifang, YANG Bo, et al. Workflow-based meteorological user environment for high performance computing[J]. Computer Engineering, 2010, 36(8):278-280

[5] 宗翔.中国气象局高性能计算环境[J].高性能计算发展与应用,2005(3):23-26
ZONG Xiang. China meteorological administration high performance computing environment[J]. Development & Application of High Performance Computing, 2005(3): 23-26

[6] 薛正华,董小社,胡雷钧,等.高性能服务器集群部署系统传输模型研究[J].计算机学报,2008,31(11):1956-1964
XUE Zhenghua, DONG Xiaoshe, HU Leijun, et al. Study on transfer model of deployment system for high performance server cluster[J]. Chinese Journal of Computers, 2008,31 (11):1956-1964

[7] 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
WANG Bin, CHANG Biao, ZHU Jiang, et al. Design and implementation of resource monitor module in meteorological computational grid platform[J]. Journal of Applied Meteorological Science, 2009, 20(5):642-648

[8] 王春虎.国家级气象高速骨干网络的系统设计[J].应用气象学报,2002,13(5):637-640
WANG Chunhu. System design of national meteorological high-speed backbone network[J]. Journal of Applied Meteorological Science, 2002, 13(5):637-640

[9] 黄倩,汪东升.远程高性能计算环境的设计与实现技术[J].清华大学学报(自然科学版),2002,42(10):1377-1380
HUANG Qian, WANG Dongsheng. Remote high performance computing environment for low reliability networks[J]. Journal of Tsinghua University(Science and Technology), 2002, 42(10):1377-1380

[10] 徐德发.上海超级计算中心网络系统简介[J].高性能计算发展与应用,2009(3):57-59
XU Defa. Network system overview of Shanghai supercomputer center[J]. Development & Application of High Performance Computing, 2009(3):57-59

Design and implementation of remote high performance computing environment in Hunan provincial meteorological bureau

ZHU Hongwu[1] YIN Xinhuai[1] LUO Dan[2] HE Wei[1] DAI Zejun[3]

1 Hunan Meteorological Information Centre, Changsha 410118

2 Hunan Meteorological Service Centre, Changsha 410118

3 Hunan Institute of Meteorological Science, Changsha 410118

Abstract Hunan Provincial Meteorological Bureau has established the first remote high performance computing environment in Hunan province, thanks for the National Supercomputing Center in Changsha. In aspects of building the remote computing environment, this paper analyzes Hunan Provincial Meteorological Bureau's requirements for high performance computing, and its calculation requirement in meteorological service. This paper focuses on the system architecture, and the technical routes and methods on establishment of a remote computing environment with a high performance computer cluster (Tianhe-1) as the computing resource. Furthermore, give consideration to both the powerful computing capabilities of the remote computing environment and the complicated requirements for the meteorological service at provincial level, we design the corresponding system process of the remote high-performance job schedule system, and the multi-level solutions to ensure the high speed and stability of remote computing environment. Finally, we run the small-scale WRF model in the remote high performance computing environment, and the results are improved obviously in calculation speed and calculation resolution, compared with that of the previously used computer cluster.

Key words remote high performance computing environment; system architecture; job schedule system