

面向设计制造的多集群高性能计算平台构建技术

吴步祺,周智力,白 剑,燕星宇

(中国运载火箭技术研究院,北京 100076)

摘要: 工业制造与工程研制对仿真分析计算需求的不断增加、仿真分析模型越来越大、计算精度要求越来越高、多学科迭代次数越来越多,对高性能计算能力要求也呈现几何级数的增长,这一切都对高性能计算资源提出了更高的要求。经过多年的建设,陆续建成了一批独立的高性能计算系统,该模式不利于资源的共享和高效利用。为此提出了一种基于多集群构建的高性能计算平台的设计方法。

关键词: 多集群;作业调度;高性能计算

中图分类号: TP38

文献标识码: A

DOI: 10.16157/j.issn.0258-7998.223479

中文引用格式: 吴步祺,周智力,白剑,等. 面向设计制造的多集群高性能计算平台构建技术[J]. 电子技术应用, 2022, 48(12): 28-32.

英文引用格式: Wu Buqi, Zhou Zhili, Bai Jian, et al. Design and manufacturing oriented multi cluster high performance computing platform building technology[J]. Application of Electronic Technique, 2022, 48(12): 28-32.

Design and manufacturing oriented multi cluster high performance computing platform building technology

Wu Buqi, Zhou Zhili, Bai Jian, Yan Xingyu

(China Academy of Launch Vehicle Technology, Beijing 100076, China)

Abstract: The demand of industrial manufacturing and engineering development for simulation analysis and calculation is increasing, the scale of simulation analysis model is larger, the accuracy of the calculation is higher, there are more and more multidisciplinary simulation, all above needs more high performance computing capability. A number of independent high performance computing cluster is not enough conducive to resource sharing and efficiency. This paper presents a design method of high performance computing platform based on multi cluster.

Key words: multi cluster; job scheduling; high performance computing

0 引言

当今,各类工业企业数字化的应用水平不断提高,应用深度和范围不断扩大,工程研制普遍采用数字化设计和分析手段进行方案设计、系统仿真验证及优化,结构、强度、流体等专业越来越多地使用数字仿真分析的方法解决各类技术问题。随着工程研制中仿真分析模型越来越大、计算精度要求越来越高、多学科迭代次数越来越多,对算力的需求呈现井喷式的增长,高性能计算系统作为可以提供超强算力的计算机,成为工程研制不可或缺的重要数字化基础条件,其已成为综合竞争力的重要标志,能强力支撑和促进工程研制,推动创新发展。多个计算能力不同的高性能计算集群间未实现互联互通,计算作业无法跨集群运行,管理复杂度高,资源效能未得到充分发挥。基于多个高性能计算集群构建逻辑统一的高性能计算平台,可以实现计算资源的共享,高效地向用户提供仿真分析服务。

迟学斌^[1]等人对高性能计算及其应用进行了综述,给出了中科院高性能计算环境建设和高性能计算应用的发展情况等,为高性能计算的发展提供了参考。任柯燕等^[2]提出了高性能计算门户的开发和应用研究中存在的若干问题,为高性能计算门户中涉及的挑战、问题及采用的技术路线提供了基础,特别是对航天等工业行业的高性能计算建设提供了指导。吴松等人^[3]则结合云计算的发展,提出了结合云计算架构的高性能计算架构的概念。在此基础上,杨学军^[4]、陈国良^[5]等人探讨了高性能计算在云环境下的新挑战、新发展,为后续基于云的高性能平台建设提供了参考。

1 设计思路

基于多个高性能计算集群分散部署于不同机房的现状,可按不同思路构建高性能计算平台。例如:各集群沿用现有调度软件和存储系统,采用分级调度方式;将所有设备集中安装于一个机房内,采用一级调度方式,整

体重新构建高性能计算平台。综合考虑现有条件、计算和存储性能及可扩展性,保持各集群部署位置不变,采用如下思路进行构建:

(1)统一门户:用户通过统一的访问门户使用高性能计算资源。

(2)统一用户:用户使用统一身份登录访问门户,提交高性能计算作业。

(3)统一调度:使用统一的调度系统,统筹调度高性能计算作业、分配资源,最大化利用高性能计算资源。

(4)统一存储:将物理分散的存储统筹整合,确保存储资源充分利用。

(5)统一监控:对包括高性能计算作业、资源利用情况、许可证等运行和使用情况进行统一监控和统计分析,科学调配计算资源,提高运维分析工作效率。

2 技术架构

2.1 设计原则

(1)高扩展性:支持多个高性能计算集群接入。

(2)高安全性:集群互联后,保证数据传输和存储的安全性。

(3)高易用性:用户可便捷高效地使用高性能计算资源。

2.2 体系架构

由管理中心和计算中心共同构成的资源集中和管理统一的高性能计算平台,实现高性能计算资源的整合与共享。

管理中心负责提供用户访问入口,用户管理、作业的统一调度和平台的监控,计算中心负责运行计算作业。

高性能计算平台主要包括如下功能模块:

(1)访问门户:用户、管理员可通过门户访问、使用和管理高性能计算平台。

(2)用户管理:通过和现有证书签发和验证系统集成,对用户证书认证;通过和主数据系统集成,获取用户和组织机构信息。计算资源通过院管理中心的活动目录进行操作系统用户的认证。

(3)作业调度:各计算中心独立建设高性能计算集群,并对各自的集群进行作业调度。实现在分布式环境下硬件资源共享,当某个计算中心资源不足时,排队等待的计算作业可以自动转发,使用其他计算中心的计算资源完成计算,相关的输入输出数据可以自动转发和返回。

(4)文件系统:采用统一的文件系统,存放各计算中心运行计算作业产生的中间数据和结果文件等。

(5)资源监控:各计算中心收集自身运行数据,管理中心汇总所有计算中心的数据,实现全局统计分析和计费。

高性能计算平台架构如图1所示。

2.3 门户设计

在管理中心部署访问门户,所有用户通过统一的访问门户使用高性能计算资源。访问门户采用集群方式部署,可实现网络负载流量均衡并满足高可用性的要求。

访问门户主要包括作业管理、数据管理、图形交互、编译调试、第三方系统集成和页面定制等功能。管理员可通过门户进行集群、作业、用户、权限和项目等的管理;用户可通过门户提交、监控管理作业和数据。

平台设计采用B/S架构,通过浏览器使用访问门户。

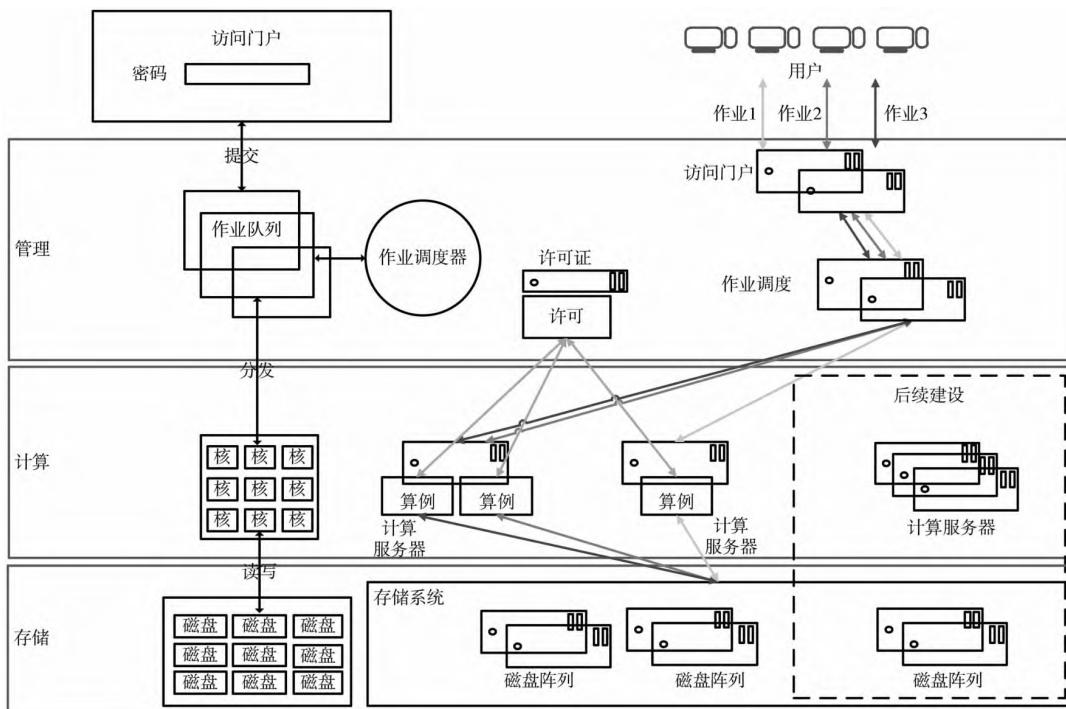


图1 多集群资源共享结构图

2.4 多集群管理与调度

2.4.1 用户和权限管理

用户身份信息由管理中心统一管理。通过和现有认证系统集成,对用户证书认证;通过和主数据系统集成,获取用户和组织机构信息。计算资源通过管理中心的 LDAP 进行操作系统用户的认证。用户管理示意图如图 2 所示。

采用分级授权的方式,平台管理员可管理所有权限,计算中心管理员仅可管理本中心的相关权限。

2.4.2 作业调度和软件管理

管理中心统一对作业进行调度。管理员可对运行作业的 CPU 时间、内存大小、运行时间等设置使用限制,对作业的优先级进行调整,并对作业进行挂起、恢复等操作。

作业调度可以配置如下调度策略:

(1)先到先算

计算作业默认按照先到先算进行排序,即计算作业的派发顺序是根据其在队列中的顺序决定的。

计算作业的顺序和优先级有关。用户或者管理员可以通过修改计算作业的优先级来改变计算作业的派发顺序。

(2)公平共享

将计算资源按用户和用户组进行分配来提供对资源的公平访问。公平共享调度能规定用户或用户组对计算资源的使用份额,保证计算资源能被公平合理地使用。

当某个用户的计算作业量不到他所占的资源份额时,其他用户的计算任务可以使用该用户份额内的剩余

资源,以保证计算资源得到最大的利用。而当该用户提交更多的计算作业时,其计算作业将比其他已使用完资源份额的用户更优先得到计算资源而执行。

公平共享调度策略可为指定的用户分配较高的使用份额,可以为指定队列配置公平共享调度策略,其他的队列使用先到先算策略。

(3)资源限制

资源限制调度策略可以限制资源的使用。如果计算作业占有超过指定数量的资源,被标记或者降低优先级。

通过设置队列参数对计算作业可使用的资源进行限制。资源限制控制了计算作业能占有多少资源。

(4)抢占调度

抢占调度策略可以使高优先级计算作业在资源紧缺时抢占低优先级计算作业,从而立刻运行。当两个计算作业竞争同一个计算资源时,高优先级的计算作业将把正在运行的低优先级计算作业挂起,高优先级计算作业立即开始运行。

当前采用先到先算结合资源限制策略对作业进行调度。默认先提交的计算作业优先级更高,优先计算,当用户到达资源限制时,将停止向用户分配资源。同时,可以对优先级进行动态调整,满足紧急计算作业需求。

仿真分析软件统一进行安装部署,如图 3 所示。

通过与访问门户和作业调度进行集成,用户提交的使用不同仿真分析软件的计算作业可运行在多个计算中心之上。

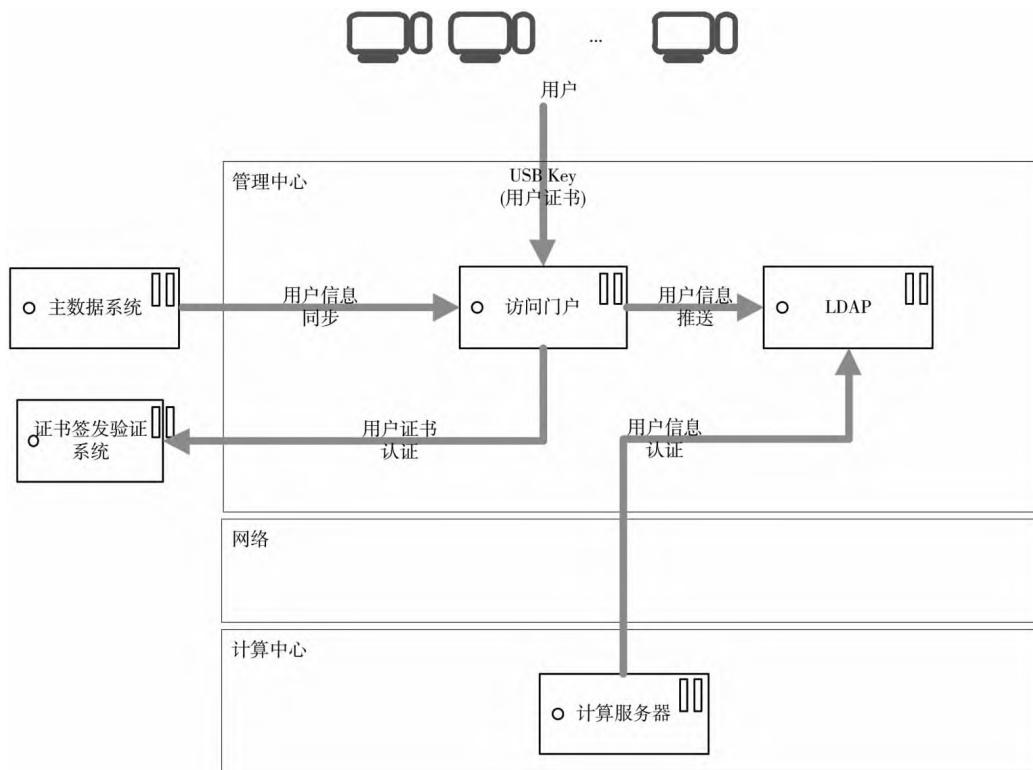


图 2 用户管理示意图

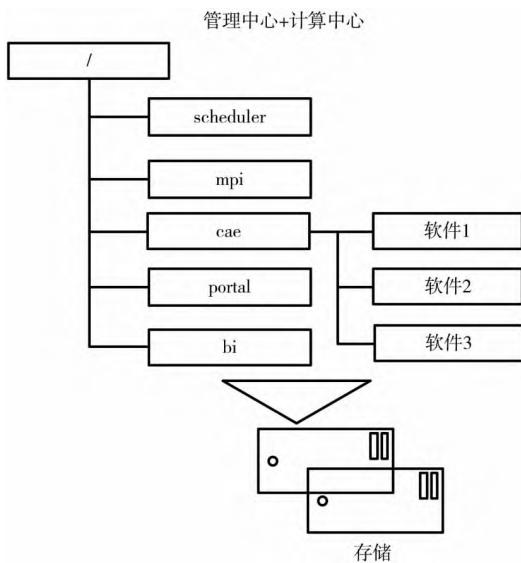


图 3 多节点仿真计算软件安装示意图

2.4.3 多计算中心调度和监控

跨集群调度示意图如图 4 所示。

(1) 计算中心互联

高性能计算平台使用的网络分为管理网、计算网和监控网。管理网用于集群的管理，计算网用于计算服务器之间的互连，监控网用于硬件的监控和控制。管理中心和计算中心通过网络互连，实现管理网、计算网和监控网的互通。管理服务器上运行作业调度服务器端程序，计算服务器上运行作业调度客户端程序，计算作业由管理中心调度到指定的计算中心运行。

(2) 文件传输

负责计算作业数据的传输，当作业准备运行时，将作业的输入文件传输至执行计算服务器的数据缓冲区目录。

通过参数配置设定是否要进行文件压缩后再传输。

文件传输模块自带断点续传功能，可避免因网络中断而导致的大文件的重传。

文件传输采用 SSL 协议，提供一种可靠的端到端的安全服务，防止客户端服务器之间的通信被攻击窃听，并且始终通过服务器进行识别认证。

使用应用层 API 对数据进行校验，保证数据的一致性和完整性。在数据传输时会对数据分块传输，最终合并成完整文件。文件传输支持断点续传功能，可避免因网络中断而导致的文件重传。

为保证作业数据安全，采用计算作业数据权限差异化管理的方式，用户仅可使用和管理其作业数据，无法查看其他用户的作业数据。

集群监控模块定期执行对所有集群的查询来获取所有集群的信息，集群监控主要包括如下功能：

(1) 实时的细粒度集群负载监控和分析：对各状态 CPU 核数占比、作业占比，CPU 核数占用率和内存利用率等进行监控和分析，图 5 所示为资源使用情况示意图。

(2) 文件系统负载和健康监控：对文件系统的空间占用情况、IOPS 等进行监控。

(3) 多维度集群负载分析：从单位等多个维度对数据进行分析。

(4) 许可证实时监控：监控许可证各个模块的使用情况。

(5) 系统日志分析和警报：针对有问题的模块进行分析和报警。

(6) 邮件报警：当出现异常时向用户发送邮件，进行提醒。

当计算作业未运行时，归入排队作业，可以从多个维度监控计算作业的排队情况。计算作业排队趋势示意图如图 6 所示。

3 结论

本文提出了一种面向工业制造的多集群高性能计

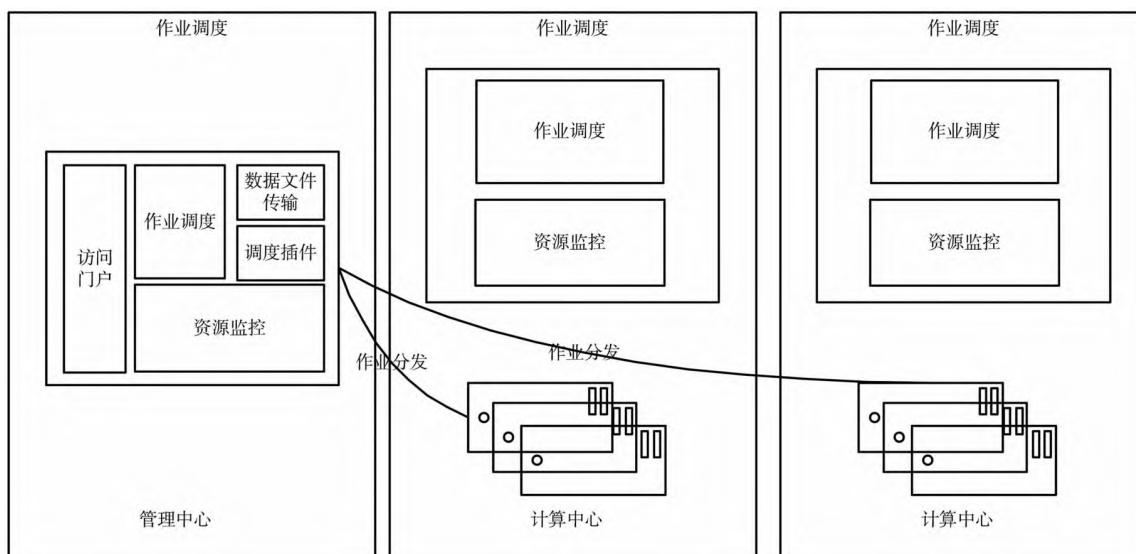


图 4 跨集群调度示意图

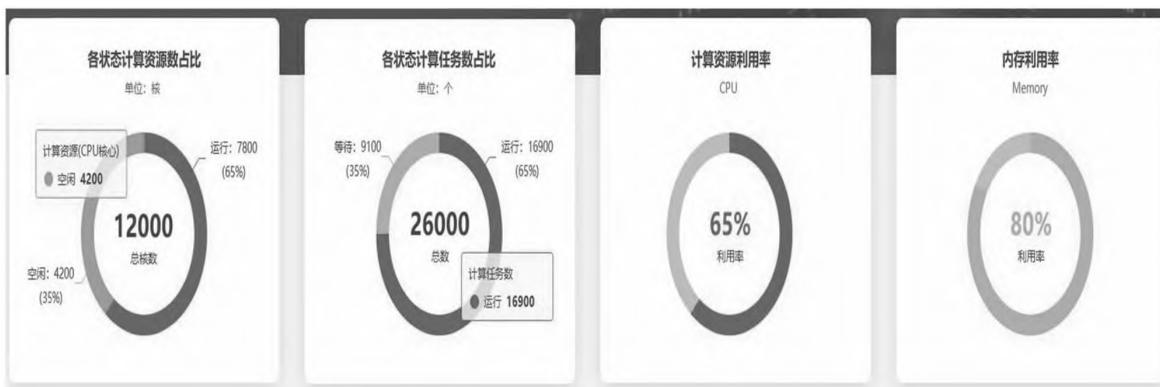


图 5 资源使用情况示意图

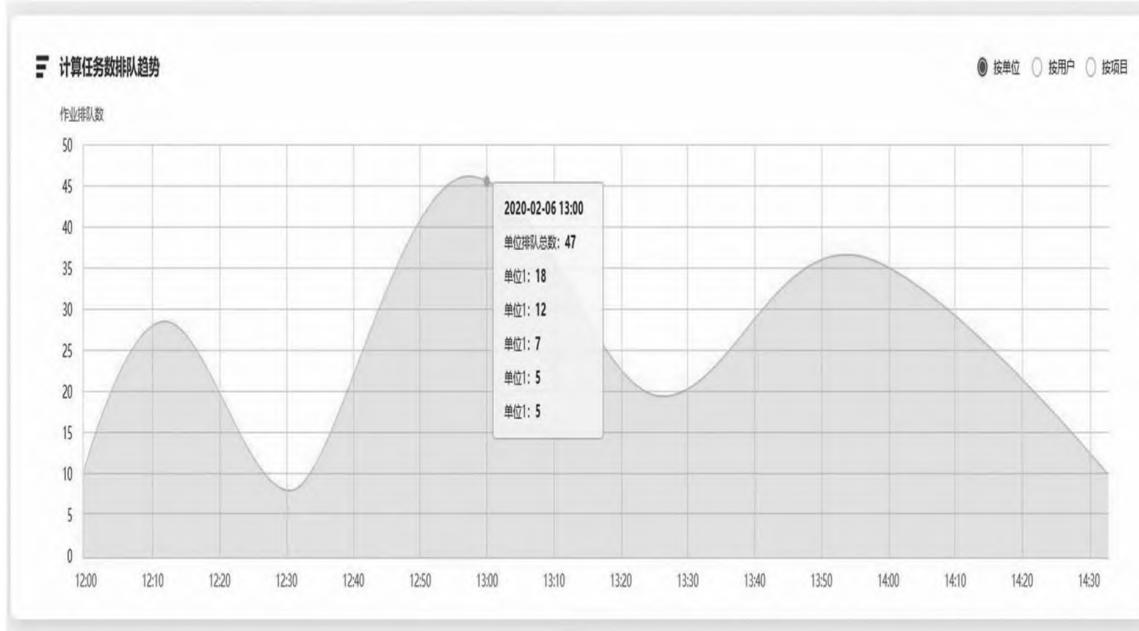


图 6 计算作业排队趋势示意图

算平台构建技术的方法。通过搭建统一访问、资源共享的高性能计算平台,提供高性能计算服务,提供软硬件支撑环境,可满足工程研制共享使用的需求,提高资源利用率,充分发挥效能。

参考文献

- [1] 迟学斌,赵毅.高性能计算技术及其应用[J].中国科学院院刊,2007,22(4):306–313.
- [2] 任柯燕.高性能计算门户开发与应用研究[J].航空科学技术,2014,25(8):69–73.
- [3] 吴松,陈海宝,金海.HPC Cloud 新兴的高性能计算模式[J].计算机学会通讯,2011,7(10):48–55.
- [4] 杨学军.并行计算六十年[J].计算机工程与科学,2012,34(8):1–10.
- [5] 陈国良,毛睿,蔡晔.高性能计算及其相关新兴技术[J].深圳大学学报(理学版),2015,32(1):25–31.
- [6] 孟柳,延建林,董景辰,等.智能制造总体架构探析[J].中国工程科学,2018,20(4):23–28.

[7] 杜娟,钱冬梅,胡东吉.智能制造在工业制造的应用[J].数字通信世界,2018,17(6):173.

[8] 吴浩,杨帆,王斌,等.基于数字孪生的火箭结构设计制造与验证技术研究[J].宇航总体技术,2021,5(2):7–13.

[9] 肖凯乐.边缘计算环境中资源分配和优化关键技术研究[D].北京:北京邮电大学,2021.

[10] 王菁.边缘计算动态调度策略的研究与应用[D].北京:华北电力大学,2021.

(收稿日期:2022-10-24)

作者简介:

吴步祺(1985-),男,研究生,高级工程师,主要研究方向:高性能计算。

周智力(1981-),男,研究生,研究员,主要研究方向:高性能计算、云计算。

白剑(1988-),男,研究生,高级工程师,主要研究方向:网络系统设计。



扫码下载电子文档