

一种采用消息模型的多集群作业管理方案设计

凌 东¹, 谷建华^{1,2}

(1. 西北工业大学 计算机学院, 陕西 西安 710072;
2. 西北工业大学 高性能计算中心, 陕西 西安 710072)

摘 要:文中针对多集群环境资源异构且地域分散、网络环境不可靠以及面向用户需求的特点,提出了一种采用消息模型的多集群作业管理方案。该方案采用全局-局部的层次调度方法,基于发布-订阅的消息模型,根据当前网络环境、用户作业的资源需求、各集群自身负载情况进行综合统一调度管理。实践证明,采用该方案设计实现的多集群作业管理系统实现了多集群环境下的资源监控、资源管理、作业调度、作业控制、数据管理等功能,有效解决了在资源异构及网络环境不可靠条件下的系统稳定性问题,显著提高了多集群系统作业吞吐能力。

关键词:消息模型;多集群;作业管理;作业调度

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2014)09-0063-05

doi:10.3969/j.issn.1673-629X.2014.09.014

Design of a Multi-cluster Job Management Scheme Using Message Model

LING Dong¹, GU Jian-hua^{1,2}

(1. College of Computer, Northwestern Polytechnical University, Xi'an 710072, China;
2. Center of High-performance Computing, Northwestern Polytechnical University,
Xi'an 710072, China)

Abstract: Aiming at the features of resources heterogeneity, geographical dispersion and unreliable network environment in multi-cluster environment, propose a method to manage job among multi-clusters using message module, which uses global-local two levels scheduling. It adopts publish-subscribe message model and can unify management and schedule job based on the present network environment, job's resource requirements and the various cluster loads. Practice has proved that the multi-cluster job management system which is designed and implemented by this method has achieved multi-cluster environment resource monitoring, resource management, job scheduling, job control, data management. It solves the stability problem effectively under the heterogeneous resources and unreliable network environment conditions. Also it significantly improves the throughput capacity of multi-cluster system.

Key words: message model; multi-cluster; job management; job schedule

0 引 言

解决大型科学和工程计算问题通常需要强大的计算能力。计算机集群通过各种互联技术将多个计算机系统连接组成一个系统,利用所有被连接系统的综合计算能力来处理大型计算问题^[1]。由于由廉价的工业标准硬件组成,计算机集群在拥有强大计算能力的同时也拥有高性价比,因而成为解决大型问题的计算机系统发展方向。

文中将主要探讨多集群环境下的作业管理方案,解决多计算集群间的作业调度问题,使用户可以通过统一平台访问和使用不同地域的计算集群资源。该平台不仅需要有效整合计算资源,提高各计算集群资源利用率和作业吞吐率,同时还要保证在网络环境不可靠的情况下稳定工作。

文中介绍了一种基于消息的多集群作业管理方案。该方案采用基于发布-订阅的消息模型,可以根

收稿日期:2013-10-27

修回日期:2014-01-27

网络出版时间:2014-05-21

基金项目:教育部项目(201104021)

作者简介:凌 东(1989-),男,硕士,研究方向为并行计算、高性能计算、异构计算;谷建华,教授,研究方向为高性能计算、分布实时数据处理、实时数据可视化、网络化嵌入式系统、计算机操作系统。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140525.1242.017.html>

据用户作业的资源需求及各个集群自身负载情况进行统一管理和调度。

1 相关工作

国内外的一些研究机构和商业公司在多集群作业管理方面已经作了相关研究和探索,但这些项目都有自己特殊的应用背景^[2-3]。下面简单介绍几个主要的类似项目。

LSF^[4] (Load Sharing Facility) 是加拿大 platform 公司的主流产品。它不仅提供单集群作业管理功能,同时还可以在单集群上安装用于多集群管理的 LSF MultiCluster 模块,实现多集群间的作业调度。LSF 具有高可用性以及强大的资源管理功能,但由于是商业软件,价格昂贵。

PBS^[5] (Portable Batch System) 是集群计算环境下出现较早的作业管理系统,主要特点有代码开放,免费获取;支持批处理、交互式作业和串行、多种并行作业等。PBS 本身主要用在单集群系统作业管理,但也提供了有限的多集群之间的作业转发机制,但是这种作业转发所基于的调度不够灵活^[6-7]。

MUSCL^[8] 是一个由英国 Warwick 开发的多集群任务负载管理系统,它提供了两级调度模式,基于 QoS 需求的动态调度算法指导作业调度。MUSCL 可以尽可能比较紧凑地在一个集群上调度作业,使用启发式算法在集群间调度,但它对任务的类型有严格限制。

2 多集群作业管理功能分析

图 1 是一个典型的多集群作业全局调度模型。

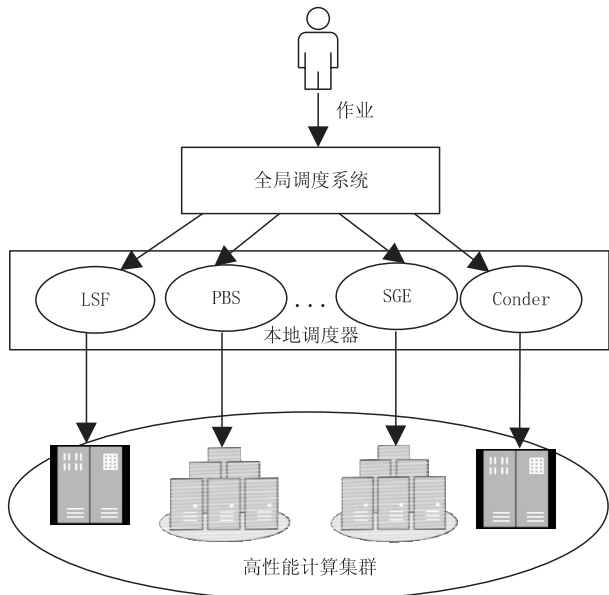


图 1 多集群作业全局调度模型

多集群作业管理在调度时首先依据全局资源状态选择某个或某些合适的集群,称为全局调度;然后作业

被分配到本地某个具体的集群,按照本地的资源管理器进行作业调度,称为本地调度;作业进入本地调度阶段后,按照本地原有的队列和调度规则进行资源分配,最终在集群各 CPU 上执行。多集群作业管理的主要功能包括:用户与权限管理、资源信息监控、作业全局调度与管理、数据传输与管理。

(1) 用户与权限管理模块。

多集群作业管理系统是运行在本地 HPC 集群作业管理系统之上的,本地 HPC 集群必然有自己的用户系统,而多集群作业管理平台也有自己的用户系统,必须在两个用户系统间设计一种用户映射的方式,同时还要设计权限控制机制。

(2) 资源信息监控模块。

资源信息监控管理是作业全局调度的基础,全局调度系统在做出决策之前,必须要事先知晓各个本地集群的 CPU 负载和内存占用比例等状态信息。因此,需要设计全局的资源监控系统,负责管理和维护系统中各个集群负载的情况,为作业管理调度决策提供支持。

(3) 全局作业调度与管理模块。

多集群调度的基本目标是协调和平衡集群间的工作负载。需要设计全局的作业管理和调度器,作业调度器根据各个集群机身负载信息做出调度决策,调度用户作业到相应的本地集群,并提交给本地集群作业管理系统。同时还要为用户设计作业状态的订阅及通知机制。

(4) 数据传输与管理模块。

由于用户作业可能被调度到任意地理位置上的集群上运行,因此,作业输入参数和计算结果需要能够在整个系统中透明的传输。同时,还要为用户提供输入参数上传和计算结果下载功能。在设计机制实现透明传输文件的同时,还必须保证该文件传输的可靠性和稳定性。

3 多集群作业管理设计思路与技术方案

3.1 消息中间件 JMS

采用 JMS^[9-10] (Java Message Service) 消息中间件来进行全局作业的调度和统一管理。JMS 的消息模型和通信特点在网络很不稳定的情况下也能保证稳定性和可靠性,并且 JMS 强大的接口能力可以方便灵活的进行定制,方便根据用户作业的资源需求及各个集群自身负载情况进行统一管理和调度。

3.1.1 JMS 特点及基本组成

JMS 可使分布式系统的通信松散连接,即发送信息的客户端只需要负责发送信息,接收信息的客户端接收信息,两个客户端之间没有必要同时可用的,甚

至发送客户端都没有必要知道接收客户端的信息,只需要发送到接收信息的服务端。

同时 JMS 还具有以下两个特征:

(1)异步的,服务端可以发送信息到一个客户端,客户端不需要为了收到信息而请求信息。

(2)可靠的,JMS API 保证了服务端所有发送的信息最少发送一次和只发送一次。

JMS 由提供者、客户、生产者、消费者、消息、队列、主题七个部分组成。其中,提供者指 JMS 的实现,可以认为是 JMS 消息服务器;JMS 客户指生产或消费消息的基于 Java 的应用程序或对象;JMS 生产者是指创建并发送消息的 JMS 客户;JMS 消费者则是接收消息的 JMS 客户;消息指可以在 JMS 客户之间传递的数据的对象;另外,JMS 队列指一个容纳那些被发送的等待阅读的消息的区域,这些消息将按照顺序发送。一旦一个消息被阅读,该消息将被从队列中移走;JMS 主题指一种支持发送消息给多个订阅者的机制。

3.1.2 JMS 的通信方式

Java 消息服务应用程序结构支持两种模型:点对点或队列模型、发布/订阅模型。

(1)在点对点或队列模型下,一个生产者向一个特定的队列发布消息,一个消费者从该队列中读取消息。在这种模式下,只有一个消费者将最终获得消息。同时,生产者不需要在接收者消费该消息期间处于运行状态,接收者也不需要在此时处于运行状态。

(2)发布者/订阅者模型支持向一个特定的消息主题发布消息。该模型如图 2 所示,对某个消息主题感兴趣的订阅者可以订阅并得到该主题的所有消息。同时,在发布者和订阅者之间存在时间依赖性。发布者需要建立一个订阅(subscription),以便客户能够购订阅。订阅者必须保持持续的活动状态以接收消息,除非订阅者建立了持久的订阅。在那种情况下,在订阅者未连接时发布的消息将在订阅者重新连接时重新发布。

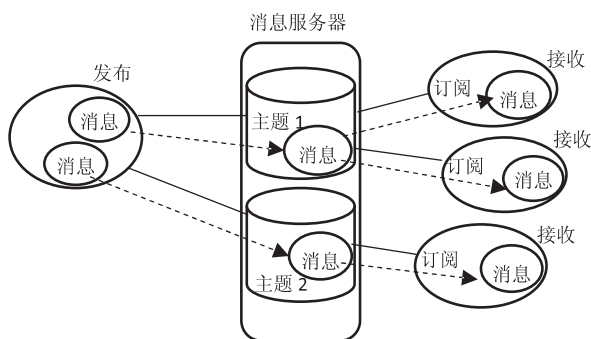


图 2 JMS 发布-订阅消息模型

3.2 本地集群作业管理

PBS、LSF 是已有常用的单个集群作业管理系统,

采用多种本地集群作业管理系统,在减少设计复杂度的同时,还可以充分满足不同本地集群管理多样性的需求。PBS 的主要特点有:代码开放,免费获取,提供完整的 API。LSF 特点是拥有强大的可用性和资源管理功能。

3.3 基于 FTP 的文件集中统一管理方案

由于多集群可能在地理位置上是分散的,而集群之间互联的网络基础设施是不可靠和不稳定的,用户需要透明地上传输入参数并下载计算结果,而不需要关心它的输入参数在哪个地方进行计算,以及需要到哪个地方去下载计算结果。

FTP 技术是比较成熟和常用的文件传输协议之一,文中采用基于 FTP 文件传输的集中统一管理方案。使用 FTP 进行文件传输,客户和服务器建立连接前要经过一个“三次握手”的过程,客户与服务器的连接是可靠的,而且是面向连接的,为数据传输提供可靠保证。它允许用户以文件操作的方式(如文件的增、删、改、查、传送等)与另一主机相互通信。

集中的文件统一管理方案可设置集中式 FTP 服务器,浏览器端提交的输入参数将传输给该 FTP 服务器,本地集群代理程序从消息服务器中获取到相应主题作业消息,解析作业消息,得到用户输入参数,到该 FTP 服务器下载相应输入参数到本地集群。本地集群代理程序查询到作业计算完成以后,将计算结果上传到该 FTP 服务器中,并把计算结果相关信息组装成消息发送到消息服务器,全局作业管理器获取到该消息,解析消息后就可以得到计算结果相关信息,此时计算结果已经在 FTP 服务器上,用户可以直接下载。

4 多集群作业管理框架与设计实现

4.1 多集群作业管理方案框架

图 3 是文中采用的基于消息模型的多集群作业管理框架图。在图 3 中,浏览器负责接收用户提交的作业描述、输入文件上传、作业状态查询、计算结果下载等功能。全局作业管理器是整个系统的核心,主要负责集中管理所有的作业,维护全局的集群系统资源负载信息,并做出相应决策,调度作业到相应的集群上运行;同时实时监控作业状态,给用户提提供作业状态查询,负责透明地将输入文件传递到相应集群,计算完成以后,再透明地将计算结果传回并透明地提供给用户下载。消息服务器主要负责缓存全局作业管理器调度给各个集群的作业信息,以及各个集群返回的作业状态信息。而各个集群的本地代理程序负责从消息服务器接收作业消息,并把消息解析成作业提交给本地作业提交系统,同时还负责定期查询本地作业管理系统该作业的状态,给消息服务器发送作业状态信息,同时

在本地计算完成以后,回传计算结果。

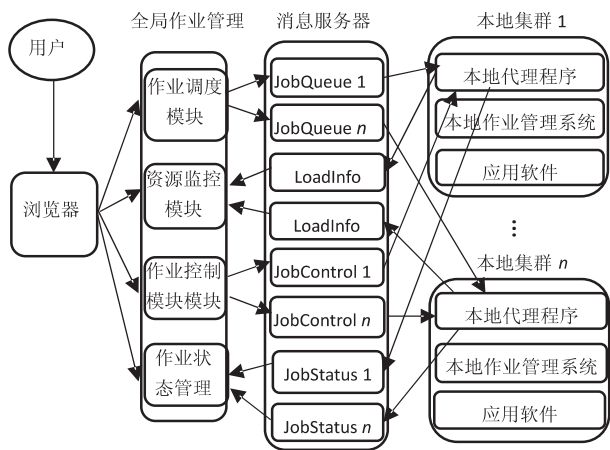


图 3 多集群作业管理框架图

4.2 多集群作业管理方案设计实现

从框架图中可以看出,该系统由六部分组成,包括作业调度模块、资源监控模块、作业控制模块、作业状态管理模块、消息服务器模块、本地集群代理程序模块。

4.2.1 消息设计

这个系统中的消息分四类,第一类是作业消息,第二类是作业状态消息,第三类是资源状态消息,最后一类是作业控制消息。

作业消息包括作业的 ID、作业脚本、作业用户名称、需求的资源、输入文件名称等。

作业状态消息即作业的状态信息,主要包括作业 ID、作业状态、作业提交时间、运行时间、完成时间等。

资源状态消息包括本地集群自身的 ID、CPU 计算能力、内存大小等静态信息,以及 CPU 利用率、主机负载等动态信息,还包括目标系统启动时间等信息。

作业控制消息包括作业 ID、作业用户名称、作业控制命令等。

4.2.2 消息服务器设计

采用的是 JMS 中发布-订阅消息服务模型作为消息服务器,类似于群发邮件的模式。消息生产者将消息发送给消息服务器,并设定一个主题,消息的消费者可以订阅其中的一个或者多个主题,并取走相应的消息。针对某个主题的订阅者,它必须创建一个订阅之后,才能消费发布者的消息,而且还可以利用 JMS 的持久化的订阅,这样,即使订阅者没有被激活,它也能接收到发布者的消息。这样就保证了在基础设施网络不稳定的情况下,消息也不会被扔掉,保证了系统的可靠性和稳定性。

消息服务器中消息主题的设计主要按本地集群编号和消息本身类型设计,及如果有 n 套本地集群,那么消息主题总数量为 $4 * n$ 。消息服务器中消息主题设

计如表 1 所示。

4.2.3 多集群系统资源监控设计

本地资源监控是多集群作业调度管理的基础,能为错误检测、资源优化配置和作业调度等提供重要的依据和参考。然而,不同本地监控系统存在着不兼容的描述或者含糊的定义,可能导致最终监控信息不准确^[11-12]。需要设计一种资源信息公共的表示方法,使得原有集群的监控信息转换成规范格式、形成一致的数据提供者。多集群系统资源监控的核心工作是本地监控信息数据采集转换及信息组织。

表 1 JMS 消息服务器消息主题设计

主题名称	内容描述
JobQueue 1	本地集群 1 作业信息
JobQueue n	本地集群 n 作业信息
LoadInfo 1	本地集群 1 负载信息
LoadInfo n	本地集群 n 负载信息
JobControl 1	本地集群 1 作业控制信息
JobControl n	本地集群 n 作业控制信息
JobStatus 1	本地集群 1 作业状态信息
JobStatus n	本地集群 n 作业状态信息

统一规范的集群监控信息包括本地集群自身 ID、CPU 频率等静态信息,以及 CPU 利用率、内存占用率等动态信息。

本地集群代理程序是守护进程,会定期向本地集群发送监控查询请求,并将返回的监控结果转换成设计的统一规范的监控信息,组装成消息,发送给消息服务器。其基本算法流程描述如下:

- (1)启动本地监控系统;
- (2)本地守护进程从本地监控系统发送监控查询请求;
- (3)本地守护进程获得返回结果以后,解析返回结果,组装成设计好的统一规范的监控信息格式;
- (4)本地守护进程将该实时监控信息发送给消息服务器;
- (5)代理程序睡眠一段时间,醒来后返回第二步。

当全局作业管理器发现消息服务器中相应主题有新消息达到,将自动获取该消息,并添加到全局资源监控数据结构中。

4.2.4 多集群作业调度模块设计

多集群的作业调度器采用基于全局-本地的两级调度机制,即全局调度和本地调度^[13-14]。每级调度都由相应的队列和调度器完成,局部调度由本地资源管理器提供,因此文中主要关注全局调度。

实现全局调度的基本手段是定义若干全局的作业队列,包括就绪队列、运行队列、完成队列。这些全局作业队列由全局调度器管理。作业最终经过全局调度

-本地调度-CPU 调度,形成了不同层次的调度。作业调度的流程设计描述如下:

(1)接收浏览器端用户提交的作业请求描述和调度说明,组装成作业对象,加入到就绪队列之中。

(2)通过监控信息系统得到系统运行状态,提供资源的使用状况及所运行作业的状态查询功能。

(3)分析作业请求描述及调度说明,匹配可用资源,得到候选资源集合。

(4)根据调度策略调度算法实现作业到特定集群资源的匹配,将作业转发到消息服务器上,同时作业从就绪队列中出队,加入到运行队列中。

(5)相应的本地守护进程从消息服务器取出作业请求描述消息,提交给本地作业调度器,并最终将作业分配到具体的处理机上运行。

4.2.5 作业控制及状态管理模块设计

作业控制、作业状态管理也是作业管理的重要组成部分。作业控制消息发送到消息服务器,相应的本地集群代理程序获取到该控制消息,并提交本地作业管理软件。

本地集群代理程序定期向本地作业管理软件发送作业状态查询请求,如果发现作业状态发生改变,将状态消息发送到消息服务器,全局作业管理器获取到该消息,解析该消息,将相应作业从运行队列中取出,加入到完成队列。

5 结束语

文中结合中国教育科研网格材料高性能计算服务门户系统项目建设背景,就如何整合不同时期、不同地域建设的高性能计算集群,提高各计算集群资源利用率,使用户可以通过提供的统一平台访问和使用这些高性能计算资源展开讨论。设计并实现了一种采用消息模型的多集群作业管理方案。该方案可以根据用户作业的资源需求及各个集群自身负载情况进行统一管理和调度。笔者基于这个原理设计并实现了该原型系统。采用该方案设计实现的多集群任务管理系统性能稳定,能实现多集群资源监控、资源管理、作业调度、作业控制、数据管理等功能。跨集群作业管理实现了不

同集群利用率均衡的同时,有效解决了在资源异构及网络环境不可靠条件下的系统稳定性问题,显著提高了多集群系统作业吞吐能力。

参考文献:

[1] Buyya R. 高性能集群计算:结构与系统(第一卷)[M]. 英文版. 北京:人民邮电出版社,2002.

[2] 张小林,钟亦平. 基于集群系统的资源管理系统的性能分析与比较[J]. 计算机应用研究,2003,20(9):56-59.

[3] 雷 州,徐志伟,祝明发. 机群管理系统的比较与评价[J]. 计算机科学,1999,26(8):23-26.

[4] LSF Team. LSF administrator's guide[EB/OL]. 2007. <http://www.platform.com/products>.

[5] Bayucan A. Portable batch system open PBS release 2.3 administrator guide[EB/OL]. 2000. <http://www.pbspro.com>.

[6] 李全枝,梁正友. 集群资源管理系统 PBS 及其应用[J]. 微机发展(现更名:计算机技术与发展),2005,15(4):4-7.

[7] 李 源,郑全录,曾 韵. PBS 作业管理系统分析[J]. 现代计算机:上下旬,2004(3):17-19.

[8] He Ligang, Jarvis S A, Spooner D P, et al. Dynamic scheduling of parallel jobs with QoS demands in multiclusters and grids [C]//Proc of fifth IEEE/ACM international workshop on grid computing. [s. l.]:[s. n.],2004:402-409.

[9] 王 军. 基于 JMS 的消息中间件设计与实现[J]. 计算机应用,2003,23(8):64-67.

[10] 朱方娥,曹宝香. 基于 JMS 的消息队列中间件的研究与实现[J]. 计算机技术与发展,2008,18(5):172-175.

[11] Schwarzkopf M, Konwinski A. Omega: flexible, scalable schedulers for large compute clusters [J]. ACM Transactions on Computer Systems, 2008, 26(4):24-26.

[12] Cao Junwei, Spooner D P, Turner J D, et al. Agent-based resource management for grid computing [C]//Proc of 2nd IEEE/ACM symposium on cluster computing and the grid. [s. l.]:IEEE,2002:350-350.

[13] Li Hui, Groep D, Wolters L. Workload characteristics of a multi-cluster supercomputer[C]//Proceedings of the 10th international conference on job scheduling strategies for parallel processing. [s. l.]:[s. n.],2004:176-193.

[14] 武 斌. 网格市场环境下资源调度机制研究[D]. 合肥:中国科学技术大学,2010.

(上接第 62 页)

[9] 周子亮. 结合非负矩阵分解的推荐算法及框架研究[D]. 北京:北京交通大学,2013.

[10] 郁 雪,李敏强. 一种结合有效降维和 K-means 聚类的协同过滤推荐模型[J]. 计算机应用研究,2009,26(10):3718-3720.

[11] 徐 红,彭 黎,郭艾寅,等. 基于用户多兴趣的协同过滤策略改进研究[J]. 计算机技术与发展,2011,21(4):73-


76.

[12] Tsi C F, Hung C. Cluster ensembles in collaborative filtering recommendation[J]. Applied Soft Computing, 2012, 12(4):1417-1425.

[13] 蔡晓妍,戴冠中,杨黎斌. 谱聚类算法综述[J]. 计算机科学,2008,35(7):14-18.

[14] von Luxburg U. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4):395-416.

一种采用消息模型的多集群作业管理方案设计

作者:	凌东, 谷建华, LING Dong , GU Jian-hua
作者单位:	凌东, LING Dong(西北工业大学 计算机学院, 陕西 西安, 710072) , 谷建华, GU Jian-hua(西北工业大学 计算机学院, 陕西 西安 710072; 西北工业大学 高性能计算中心, 陕西 西安 710072)
刊名:	计算机技术与发展 
英文刊名:	Computer Technology and Development
年, 卷(期):	2014(9)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjtz201409014.aspx