

# 集群作业管理系统的关键技术分析比较

段新华,王宏勇,丁 汨

(河南工业大学 信息科学与工程学院,河南 郑州 450001)

**摘要:**集群技术是一种较新的技术,通过集群技术,可以在付出较低成本的情况下获得在性能、可靠性、灵活性方面的相对较高的收益,集群作业管理系统则是集群系统中的核心部分。讨论了集群作业管理系统中的一些关键技术,如作业调度的组织模式、作业调度策略、进程迁移机制、资源组织和管理等,据此分析比较了当今具有代表性的几种集群作业管理系统 PBS、CONDOR 和 LSF,并由此得出了对于今后研究和开发下一代集群作业管理系统有重要指导意义的结论。

**关键词:**集群系统;作业管理系统;技术分析

中图分类号:TP391.41

文献标识码:A

文章编号:1673-629X(2009)02-0087-04

## A Core Technology Analysis and Comparison for Cluster Job Management System

DUAN Xin-hua, WANG Hong-yong, DING Mi

(School of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China)

**Abstract:** Cluster is a kind of newer technology, through which the whole network, under lower costs, can obtain higher benefits on reliability and flexibility, and cluster job management system is the core technology, this paper emphatically discusses some key technologies used in job management system, such as job scheduling organization pattern, job scheduling policy, process migration mechanism, resource management, then introduces several prevailing and representative cluster job management system, including PBS, CONDOR and LSF.

**Key words:** cluster system; job management system; technology analysis

### 0 引言

集群是一种并行或分布式处理系统,由很多连接在一起的独立计算机(称为节点)组成,各节点协同工作,就像单个集成的计算资源一样,计算机节点可以是一个单处理器或多处理器的系统(PC、工作站或SMP),拥有内存、I/O设备和操作系统<sup>[1]</sup>。集群提供了单个系统不能支持的高可用性、高可靠性和易伸缩性,以其卓越的性能价格比和良好的扩展性成了当今高性能计算的主流体系结构。集群的作业管理系统JMS(Job Management System)是集群系统软件的重要组成部分,是保证集群高效使用的关键,它可以根据用户的需求,统一管理和调度集群系统的软、硬件资源,保证用户公平合理地共享资源,形成对用户透明的单一管理系统,提高资源的利用率和吞吐率,这使得集群作业管理系统一直成为集群计算环境的研究热点<sup>[2]</sup>。

文中通过分析集群作业管理系统的关键技术,比较了当今具有代表性的一些集群作业管理系统,并由此得出了一些对于今后研制集群作业管理系统具有重要指导意义的结论。

### 1 集群作业管理系统的关键技术

根据作业管理系统的两大功能,即资源组织管理和作业调度,分析集群作业管理系统的关键技术。

#### (1) 作业调度的组织模式。

集群作业管理系统中作业调度常用的组织模式有集中式、分布式、层次式<sup>[3]</sup>。

①集中式组织模式:集群作业管理系统中只有一个作业调度器,该作业调度器保存全局资源视图,并根据系统资源的变动情况实时更新资源视图。作业调度器接收该系统中所有作业请求,统一进行调度。集中式组织模式的优点在于由于信息的集中式收集和保存,便于管理。其缺点在于在系统规模扩大时,作业调度器往往成为性能和管理的瓶颈。

②分布式组织模式:在每个参与调度的计算节点上驻留一个作业调度器,接收用户的作业信息,当其资

收稿日期:2008-05-20

基金项目:河南省重大科技攻关基金资助项目(072SGZS38042)

作者简介:段新华(1980-),女,河南濮阳人,硕士研究生,研究方向为高性能计算、集群系统结构;王宏勇,博士,副教授,硕士生导师,研究方向为高性能计算、图像处理。

源信息发生变化时,向其他节点广播,从而保持系统内资源视图的一致性。分布式组织模式中不存在一个集中的作业调度器,避免了集中式作业调度器成为系统性能瓶颈的弊端,但在各个调度期间需要保持资源视图的一致,需要进行频繁的广播通讯,从而影响了分布式组织结构的扩展性。

③层次式组织模式:层次式组织模式是集中式和分布式组织结构的折衷。在这种组织模式下,将集群计算环境分为几个独立的调度域,在每个调度域中设置一个作业调度器,在单个调度域中无法解决的作业调度可向其上层作业调度器报告,上层作业调度器中保存跨域的资源视图,对报告的作业请求进行调度。层次式组织模式减少了调度其成为性能瓶颈的可能性,并具有较好地扩展性。层次式调度的组织模式典型的例子就是集群环境下的分区技术,通过将集群的资源进行分区,较好地适应了集群的应用多样化和策略多样化的需求。

#### (2)作业调度策略。

集群作业管理系统的作业调度策略可分为作业选取策略和资源分配策略<sup>[4,5]</sup>。

作业选取策略是指集群作业管理系统以何种规则从作业队列中选取待执行作业,资源分配策略是指集群作业管理系统根据执行作业的资源需求,如何为作业分配合适的资源使其投入运行。

作业选取策略可分为面向公平性的策略和面向高效率的策略两类。面向公平性的作业选取策略主要有先来先服务(FCFS)策略、短作业优先(SPF)策略等。面向高效率的作业选取策略主要有 FirstFit 策略、BestFit 策略等。

资源分配策略可分为时间共享和空间共享两类。时间共享的主要特征是作业调度器周期性地为作业分配资源,在下一个周期,不管运行在资源上的作业是否结束,该作业都有可能被终止或挂起,作业调度器有可能将资源分配给其他作业使用。空间共享的主要特征是作业一旦分配到资源就投入运行,直到作业执行结束。

空间共享资源分配策略通常分为以下三类:

①资源独占策略:资源在任何时刻只能分配给一个作业使用。当资源已分配给一个作业后,只要该作业还未完成,它就不能分配给其它作业。资源独占策略对于提高并行作业的运行效率,缩短运行时间具有重要意义。但是,在资源独占策略中,资源通常不能满负荷运行,不利于提高系统资源的利用率和吞吐率,影响整个系统性能的提高。

②公平共享策略:系统中的每个资源都可同时运

行多个作业,在分配时,根据系统中各资源的利用率、负载状态和的计算能力以及作业的资源要求来进行资源的选取。当资源上已经有作业在运行时,只要其负载未超过一定的阈值,系统仍可以将作业分配到该资源上执行。在一个资源上运行的多个作业,根据公平共享原则,由操作系统对其进行低级调度,即进程调度。公平共享策略可以使系统满负荷运行,对于提高系统的利用率和吞吐率具有重要的作用。公平共享是目前集群资源分配中广泛使用的策略。

③可变分配策略:是针对并行作业采用的一类资源分配策略。在该策略下,用户作业周期性地向资源分配器提供其在下一个周期所需要的资源数量,资源分配器在每个分配周期进行资源的重新分配,将资源“过剩”的作业所使用的部分资源,分配给资源“短缺”的作业,从而提高资源的利用率。可变分配是一类理想的资源分配策略,但在实现上,对用户作业提供信息的要求较高,需要用户使用系统提供的特定语言接口编程,以支持用户周期性的资源需求信息收集。

#### (3)进程迁移机制。

进程迁移是指将一个正在运行的作业进程从集群系统的一个计算节点迁移到另一个目标计算节点上,迁移后的作业进程能从迁移点处继续往下执行,其行为和结果与没有发生迁移一样。成熟的进程迁移机制应具备的主要特征是透明性、无剩余依赖、一致性和高效性,其中透明性是对进程迁移机制的首要要求,体现在对用户的透明性、对系统的透明性和对进程本身的透明性<sup>[6]</sup>。

进程迁移可根据应用的级别分为用户级迁移、应用级迁移和系统级迁移三类<sup>[7,8]</sup>:

①用户级迁移:用户级迁移实现较为简单,软件开发和维护也较为容易,因此,现有的很多系统都是采用用户级实现,如 Condor 和 Utopia。但由于在用户级无法获得 Kernel 的所有状态,因此,对于某类进程,无法进行迁移。另外,由于 Kernel 空间和 User 空间之间存在着壁垒,打破这个边界获得 Kernel 提供的服务需要巨大的开销。因此,用户级迁移的效率远远低于内核级迁移的实现。

②应用级迁移:应用级迁移的实现较为简单,可移植性好,但是需要了解应用程序语义并可能需对应用程序进行修改或重新编译,透明性较差,这方面的系统有 Freedman, Skordos 等。

③内核级迁移:基于内核的实现可以充分利用 OS 提供的功能,全面的获取进程和 OS 状态,因此实现效率较高,能够为用户提供很好的透明性。但是由于需要对 OS 进行修改,实现较为复杂。这方面的典型系

统有 MOSIX 和 Sprite 系统。

#### (4) 资源组织和管理。

集群作业管理系统要为系统资源提供有效的描述手段,分配和监控系统资源,收集资源信息,并且对资源进行合理的组织,才能有效地进行作业调度,有效执行各类应用。这里的资源是个很广泛的概念,各种硬件设备、数据和程序都可以看成资源,如 CPU、存储、网卡,甚至系统的事件和 logo。

资源组织和管理的关键技术主要包括资源模型、资源名字空间组织、资源信息的存储和资源发现。在集群环境下,由于资源相对的同构和稳定,作业管理系统的研究重点在于作业调度模式的选择<sup>[9]</sup>。

## 2 对各种集群作业管理系统的分析和比较

目前已有的集群作业管理系统在目标、结构、功能和实现上各有差异,从不同侧面反映了集群作业管理系统所应具备的特性。PBS、CONDOR、LSF 是当今颇具代表性和影响力的几种集群作业管理系统。其中 PBS、CONDOR 是研究产品,LSF 是商业软件。本节对这几种作业管理系统进行简单介绍和讨论<sup>[10,11]</sup>。

### 2.1 PBS(Portable Batch System)

PBS 最初由 NASA 的 Ames 研究中心开发,为了提供一个能满足异构计算网络需要的软件包,特别是满足高性能计算的需要<sup>[12]</sup>。它力求提供对批处理的初始化和调度执行的控制,允许作业在不同主机间的路由。PBS 的独立的调度模块允许系统管理员定义资源和每个作业可使用的数量。调度模块存有各个可用的排队作业、运行作业和系统资源使用状况信息。PBS 由四个组件组成:命令组件 Commands,作业服务器 Job Server,作业调度器 Job Scheduler、作业执行器 Job Executor。PBS 作为开放源码的软件包,被广泛利用。用户可根据自己的需要对源码进行相应得修改,因此,不同的系统通过对 PBS 的改造能够形成面向具体系统的作业调度系统<sup>[13,14]</sup>。

(1)在作业调度的组织模式上,PBS 采用了集中式组织模式。PBS 本身不支持分区概念,但使用者可以基于 PBS 构建多分区调度系统,在每个分区中采用 PBS 进行作业调度。

(2)在作业调度策略上,PBS 提供了默认的 FIFO 调度策略,还提供了 TCL,BACL(Batch Scheduling Language),C 三种过程语言和调度类,并定义了一些调度需要的函数,以及提供了完整的 API,用户只需修改调度类就可方便实现新的调度策略。PBS 实现了优先级策略,并且其作业调度和节点分配策略是可配置的,对于节点分配策略,提供了公平共享和独占两种策略。

除了本身的调度外,PBS 还可以使用 maui 调度。

(3)PBS 本身未提供检查点操作和进程迁移,但在操作系统支援的情况下,可支持检查点的操作。PBS 实现了自动的负载平衡,提供了内存使用率、负载均值等多个负载参数。

(4)在 PBS 系统中,系统管理员能控制资源的可用性,能更改资源配置,添加、删除和修改资源,能控制用户对资源的存取权限。用户提交作业时能确定资源要求,提交作业后能更改作业的资源要求。

### 2.2 CONDOR

CONDOR 是由 Wisconsin - Madison 大学计算机系开发的。其目标是通过开发网络上的当前计算资源来为用户提供高吞吐量的计算环境。充分利用工作站的空闲时间是 CONDOR 的最显著特征。CONDOR 管理的集群由网络中的工作站组成。工作站主人可以自愿加入或退出。CONDOR 监测网络中所有工作站的状态,一旦某台计算机被认为空闲,便把它纳入到资源池(Pool)中。在资源池中的工作站被用来执行作业。在 CONDOR 中,每个用户都由一个用户代理代表,每个资源由一个资源代理来代表<sup>[15,16]</sup>。

(1)CONDOR 的作业调度组织模式经历了从集中式到层次式的发展,其层次式的组织模式在引入分区的概念后,通过匹配器实现了全局调度器的功能。

(2)CONDOR 在作业调度策略上,采用了 FCFS 和基于优先级的作业选取策略,公平共享的资源分配策略,CONDOR 着重于资源的匹配,提供了一套称为 Match making 的资源匹配机制。CONDOR 只有本身的调度策略。

(3)CONDOR 实现了应用级的进程迁移,但不支持并行作业的迁移,对于标准作业,可以比较全面地实现检查点的操作。CONDOR 定期地对作业的执行设置检查点,甚至是在作业被迁移时,也会设置检查点,目的是为了保留作业在因意外(系统崩溃,突然掉电等)被丢失先前已经使用的资源现场。

(4)在 CONDOR 系统中,支持各个计算节点的动态加入和退出,工作站的拥有者对该节点有完全的控制权,能指定节点资源的可用性和可用时间,系统能根据键盘或鼠标空闲时间或网络状态决定该计算节点是否加入<sup>[17]</sup>。

### 2.3 LSF(Load Sharing Facility)

负载共享软件 LSF(Load Share Facility)是由加拿大平台计算(Platform Inc.)公司研制与开发的,由 Toronto 大学开发的 Utopia 系统发展而来。从强大的功能和广泛使用的角度看,LSF 可谓是一个成熟的集群作业调度系统。在使用的范围上,LSF 不仅用于科

学计算,也可用于企业的事务处理。功能上,除了一般的作业管理特性外,它还在负载均衡、系统容错、检查点操作、进程迁移等方面作了很好的努力,并力图使之实用化<sup>[18]</sup>。

(1)在作业调度的组织模式上,LSF 支持层次式组织模式。

(2)LSF 的作业调度模式提供了可扩展的作业选取策略框架,支持多种作业选取策略,并允许用户自行确定策略,并提供了抢占式调度和关键资源保障,保证紧急作业的调度。LSF 在资源分配上提供公平共享和独占式策略。

(3)LSF 支持核心级、用户级及应用程序级的进程迁移和检查点操作。

通过以上对各个集群作业管理系统在关键技术上的介绍和比较,可以看到各个系统在实现和功能上大同小异。PBS 作为一种应用范围较广的开放源代码的自由软件,有着良好的性能和广阔的市场发展前景。可以对 PBS 的作业调度策略进行扩展,取代其默认的 FIFO,这样可以使集群作业管理系统有更优秀的调度能力,用户就可以运行更多的作业和更快的得到运行结果。CONDOR 是一个功能很强的研究产品,在工作站网络中适当使用,工作站的空闲处理能力将以更高的效率得以发挥,缩短了作业响应时间,并在检查点操作和进程迁移方面做了很大的努力,是想充分利用局域网内工作站计算资源的系统管理员的最好选择<sup>[19]</sup>。此外,CONDOR 仅支持单进程的任务,不支持用户应用程序使用进程间通信,其对于任务的局限性是由于系统设计造成的,改进难度较大。LSF 可以满足商业和技术两个领域的工作管理要求,是一个很成功的商业集群资源管理系统,在功能及可用性方面做了较大的努力,具有较高的实用性和优势,已成为集群管理软件领域中的国际工业标准。

### 3 结束语

一个功能强大、富有效率的集群作业管理系统不但能大大方便用户的使用,而且能够极大地提高集群系统的利用率。通过分析集群作业管理系统的核心技术,对目前几种流行的作业管理系统进行了分析比较。目前的集群作业管理系统中,提供了多种作业调度策略供用户选择,但这些调度策略多是针对集群单管理域,应用类型单一的情况的,在集群多管理域、多应用类型及异构环境下,对如何整合全局资源,进行作业调度的讨论较少。因此,在以后的研究中,如何在集群作业调度策略的设计上兼顾公平性和高效性、如何在异构集群系统间有效利用资源以实现高效作业调度,将

是研究的重点。同时,面对网格系统的出现,集群作业管理系统应在支持网格应用方面做一些研究,以适应新的计算环境。

### 参考文献:

- [1] Buyya R. 高性能集群计算:结构与系统(第一卷)(英文版)[M]. 北京:人民邮电出版社,2002.
- [2] 汤小春. 基于集群技术的作业管理系统的研究与实现[D]. 西安:西北工业大学,2001.
- [3] Li K. Performance Evaluation of job scheduling an processor allocation algorithms for grid computing on meta computers [C]//Proceeding of Parallel and Distributed Processing Symposium. [s.l.]:[s.n.],2004:170-175.
- [4] Karatza K H. A simulation model of task cluster scheduling in distributed system[C]//Proceedings 7th IEEE Workshop on Future Trends. [s.l.]:[s.n.],1999:163-168.
- [5] 叶庆华,孟丹,江滢.曙光 3000 机群作业管理系统 JOSS 的设计与实现[J]. 计算机工程,2003,29(6):42-44.
- [6] Yi-Ming Wang. Hang Yen nun, Check pointing and its applications[C]//Twenty - Fifth International Symposium on Fault-Tolerant Computing. [s.l.]:[s.n.],2003:27-30.
- [7] 张怡,胡凯,胡建平.群机系统中进程迁移实现机制的研究[J]. 计算机工程与应用,2001(1):31-35.
- [8] 黄翊,蒋江,张民选. MOSIX 进程迁移机制研究[J]. 计算机工程,2002,28(8):117-119.
- [9] 雷州. 机群作业管理系统研究[D]. 北京:中科院计算技术研究所,1999.
- [10] 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(9):56-59.
- [11] 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.
- [12] Bayucan A. Portable Batch System OpenPBS Release 2.3 Administrator Guide[EB/OL]. 2000. Veridian Information Solution, <http://www.pbspro.com>.
- [13] 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.
- [14] 李源,郑全录,曾韵. PBS 作业管理系统分析[J]. 现代计算机,2004(3):17-19.
- [15] Basney J, Livny M. Deploying a High Throughput Computing Cluster, High Performance Cluster Computing[M]. Englewood Cliffs:Prentice Hall,1999.
- [16] Condor Team. Condor - high Throughput Computing[EB/OL]. 2008. <http://www.cs.wisc.edu/condor/>.
- [17] 郭雷,鞠九滨,张怡颖. 对 Condor 系统的分析与改进[J]. 软件学报,1997,8(5):345-349.
- [18] LSF Team. LSF Administrator's Guide[EB/OL]. 2007. <http://www.platform.com/products>.
- [19] 郭绍忠,黄永忠,余丽琼. 机群作业管理系统 Condor 综述[J]. 信息工程大学学报,2004,5(1):73-76.