第34卷 第13期 计 算 机 工 程 2008年7月
Vol.34 No.13 Computer Engineering July 2008
·软件技术与数据库· 文章编号：1000—3428(2008)13—0075—03 文献标识码：A 中图分类号：TP311

多集群并行作业的性能监测及分析

陈诗然，胡 凯，张 伟，张 璐
(北京航空航天大学计算机学院，北京 100083)

摘 要：介绍一种多集群计算模式，在分析了多集群系统结构灵活、具有可重组性等特点的基础上，研究适用于该模式的并行作业性能监测分析方法与技术，设计和实现了一个并行作业性能监测分析工具。它采用动态性能分析方法，遵循分布式软件设计架构，具有高内聚、低耦合的模块组织结构，运行验证表明其能够在多集群计算模式下有效工作。

关键词：多集群；并行作业；性能监测；性能分析；插桩

Performance Monitoring and Analysis on Multi-cluster Parallel Jobs

CHEN Shi-ran, HU Kai, ZHANG Wei, ZHANG Lu
(School of Computer Science and Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100083)

【Abstract】 A multi-cluster computing model is introduced. In the analysis on the basis of the multi-cluster system features, which include flexible architecture and reconfigurability, the methods and technologies of performance monitoring and analysis, which is applicable to this model, are researched. A performance monitoring and analysis tool of parallel jobs is designed and implemented. In this tool, dynamic performance analysis method is used, distributed software design framework is followed, a high cohesion and low coupling structure is designed. Operation results show it can work effectively in the multi-cluster computing model.

【Key words】 multi-cluster; parallel jobs; performance monitoring; performance analysis; instrumentation

多集群计算是集群并行计算发展的一种必然趋势，它能有效地凝聚计算资源，形成协同的高性能计算能力，并具有良好的可伸缩性。在多集群计算模式下，并行作业的性能监测分析是提高计算效能的一个重要手段，也是并行计算中的一个研究难点，它能够监测复杂并行作业的执行过程，揭示并行作业的执行行为，分析影响执行效率的因素和瓶颈，为改进和优化并行作业提供依据。

1 研究现状

国外在并行作业性能监测分析领域的研究非常活跃，产生了一些具有代表性的成果：(1)MPE(Multi Processing Environment)是MPICH自带的MPI并行作业性能监测库，用特定格式的日志文件存放数据，采用事后分析的方法。(2)Paradyn是Wisconsin-Madison大学开发的性能监测分析工具。它支持异构环境，使用DyninstAPI动态性能监测库对运行中的并行程序二进制映像进行插桩，实时分析，利用子图折叠算法[1]优化分析结果。(3)SCALEA是Innsbruck大学开发的自动化性能诊断工具[2]，支持MPI, OpenMP和HPF并行作业，在编译期对代码区域插桩，实时分析，并支持多实验分析；经过重新封装为SCALEA-G还可支持网格计算环境。(4)Intel公司的Trace Analyzer and Collector是商业产品，专注于MPI并行作业分析，Collector搜集性能数据产生跟踪文件，Analyzer利用跟踪文件进行分析。

国内的研究工作主要集中于各高校，如清华大学的性能分析工具 THPT ii，北京大学构筑在 HPF 编译系统之上的性能监测分析工具，国防科技大学用于银河巨型机的 YHPROF 工具等。它们都是适应自身系统特点的工具。

上述工具的共同特点是专为单集群系统或大型机结构设计，它们大多采用集中式控制方式，只能由一台节点充当全局管理节点，未针对多集群环境进行设计，历史分析数据的利用率不高。因此，有必要对多集群计算环境下的并行作业性能监测分析开展进一步的研究。

2 基于多集群的性能监测分析

2.1 多集群计算模式

多集群系统致力于把若干分布式独立集群通过专用或通用高速网络互连，利用中间件技术形成一种对用户而言透明、单一的计算环境。在实际应用中，还可能面临各种灵活的需求，比如将几个集群组合成更大规模的集群，或将多集群拆分成独立集群便于车载移动使用，提高集群的利用率和灵活性，并提供一定的高可用性。

由于多集群系统的应用环境存在差异，因此需要设计不同的计算模式。

本文研究了一种具有分布式可重组特点的多集群计算模式。该模式下的多集群环境由多个单元集群组成，各集群有独立的管理域并可单独工作，集群之间是对等工作模式，集群的组合和拆分无需改变集群的管理域，能够灵活方便地重组，具有较强的可伸缩性和高可用性，集群重组基本不影响系统运行的并行计算任务。图 1 描述了该模式下的多集群组织结构。

假设多集群由n个集群$C_1,C_2,\cdots,C_n$组成，其中，C_i(1≤ i

作者简介：陈诗然(1979－)，男，硕士研究生，主研方向：分布式并行计算；胡 凯，副教授、博士；张 伟，硕士；张 璐，硕士研究生

收稿日期：2007-08-16 **E-mail：**csr007@163.com

≤n)由m_i个同构的计算资源构成，提交给多集群的并行作业可以在任何一个有权限的集群上执行。其可重组性表现在：多集群$M=\{C_i,\ 1\le i\le n\}$中的集群可以任意组合为一个集群组$N_1,N_2,\cdots,N_m$。其中，$1<m\le n$，且$N_j(1\le j\le m)$是M的非空子集之一，如果$x\in N_p,\ y\in N_q$并且$p\ne q$，那么必然有$x\ne y$。本文采用P2P对等计算中的Chord环[3]系统来实现符合多集群计算模式的组织结构，具备结构松散、完全分布式的特点，有很好的鲁棒性，可消耗更少的资源实现可重组特性。

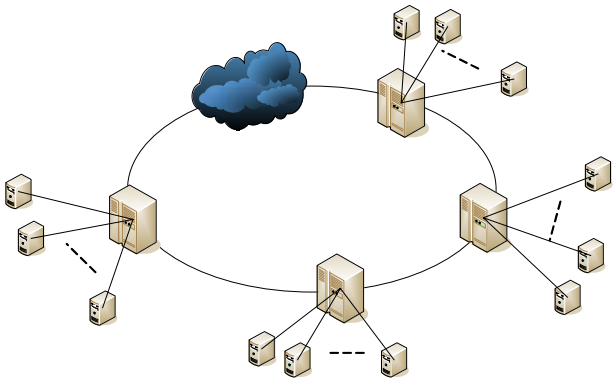

图 1　多集群计算模式组织结构

2.2　性能监测分析方法

性能分析一般分为静态性能分析和动态性能分析 2 种。静态性能分析在源程序级别进行，通过模拟程序执行或对程序分析来预测性能数据，也被称为性能预测。由于该方法并不真正运行程序，并且影响程序执行的因素有很多，因此静态性能分析无法对程序的执行情况进行完全准确的模拟，属于一种试探性的方法。

多集群计算环境具有结构灵活、可重组的特点，成员集群处于对等地位，单个或一组集群可以动态加入或移出。这就要求在该环境中进行的性能分析也必须具备相应的灵活性、实时性和高效性。本文重点研究了动态性能分析方法，该方法无需对原有并行作业进行代码修改或重新编译，而是在并行作业执行过程中采用插桩监测的方式获取性能数据，实时分析并保存分析数据。由于性能数据直接从并行作业的执行过程中获得，因此更加完整、准确，分析结果对优化和改进并行作业更有参考价值。整个过程分为如下几个主要阶段：插桩(Instrumentation)、监测(Monitoring)、分析(Analysis)与可视化(Visualization)。

插桩是为了标记和识别程序代码段中的事件触发点，位置一般在代码段的入口和出口处，插桩的对象是执行中的并行作业二进制代码，插桩行为通过工具自动进行。监测是获取并行作业性能数据的重要手段。实时分析是在并行作业执行的同时分析性能数据，并保存分析数据供日后使用。可视化则是将分析结果显示出来，为改进和优化并行作业提供依据。

2.3　实现技术

要想突破已有性能监测分析工具的设计思路，设计出符合多集群计算模式的并行作业性能监测分析方案，必须借鉴分布式系统的软件设计架构，采用高内聚、低耦合的模块组织结构，以软件系统的灵活实现来适应硬件系统的灵活结构。为了保证性能分析的实时性和准确性，尽可能减少对原并行作业的干扰，应采用动态性能监测方法，通过调用DyninstAPI动态性能监测库[4]的方式对作业进程自动插桩；同时监测系统环境信息[5]和并行作业执行信息 2 类性能数据，如CPU利用率、并行作业中的I/O通信时间、同步等待时间等；利用数据库技术存放和检索性能数据，达到提高数据存取效率、保障安全性的目的；为使产生的分析结果更具有参考价值，采用聚类分析和阈值分析等多种手段对性能数据进行实时分析，并以图表形式直观地呈现出分析结果。

3　设计与实现

3.1　系统结构

工具划分为几个模块(由于模块间的耦合度较低，也称为子系统)：底层通信库，负责封装消息通信需要的功能，屏蔽平台异构性；命令控制系统，供用户发出监测命令、控制监测行为；中心控制系统，是系统的调度核心，转发各个子系统之间的交互消息；数据采集系统，负责获取并行作业性能数据和系统环境数据；数据存储系统，通过封装简化对性能数据库的操作；数据共享系统，实现 Chord 环系统，发布实验信息；数据展现系统，提供分析结果。

系统整体层次结构如图 2 所示，大实线方框代表集群，小虚线方框代表计算节点，其余的小实线方框代表各子系统，箭头代表子系统之间的信息交互。

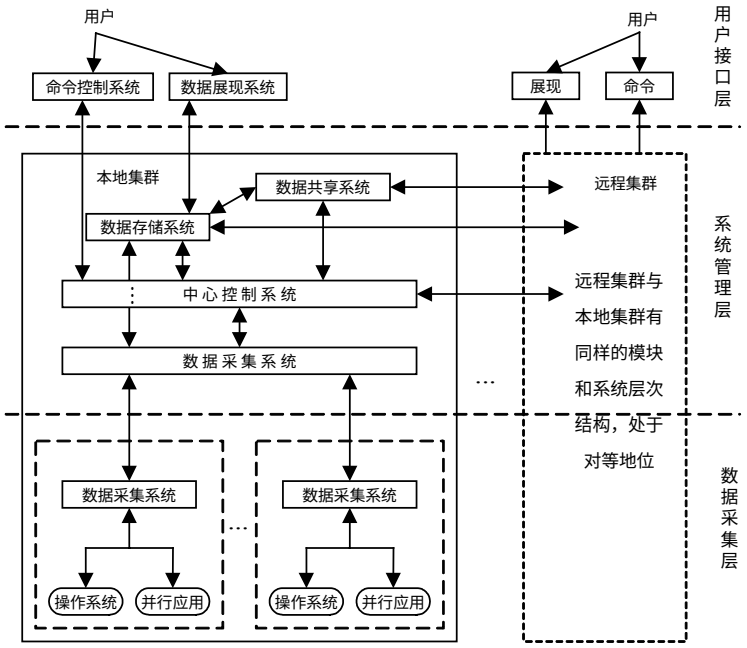

图 2　系统层次结构

从图 2 中可以看到，性能监测分析工具划分为 3 个层次：最上方为用户接口层，命令控制系统和数据展现系统工作在这一层上，该层提供了工具与用户交互的界面，可运行在管理节点或客户机上；中间为系统管理层，运行于集群管理节点，作为控制核心提供管理功能，工作在这一层上的有中心控制系统、数据存储系统、数据共享系统和数据采集系统的管理端；最下方是数据采集层，负责获取并行作业的性能数据，只包括数据采集系统的采集端，运行在集群计算节点上。

3.2　工作方式

为了降低各子系统之间的耦合度，采用了消息通信的方式，利用底层通信库来提供这一功能，中心控制系统作为消息转发的核心。子系统间的消息流和数据流如图 3 所示。

进行监测分析时，用户通过命令控制系统向中心控制系统发送监测消息；中心控制系统将监测消息转发给其他子系统；数据采集系统产生一个专门的进程负责采集该并行作业的性能数据，并将数据发送到数据存储系统；数据共享系统把本次作业实验信息发布到 Chord 环中；数据展现系统获取性能数据，显示分析结果。查询并行作业的实验数据时，一旦数据共享系统发现远程集群上有本地所需的历史数据，就向远程集群的数据共享系统发送消息，使其通知远程数据存储系统与本地交互，将历史数据发送到本地集群，并通过本地数据展现系统显示给用户。

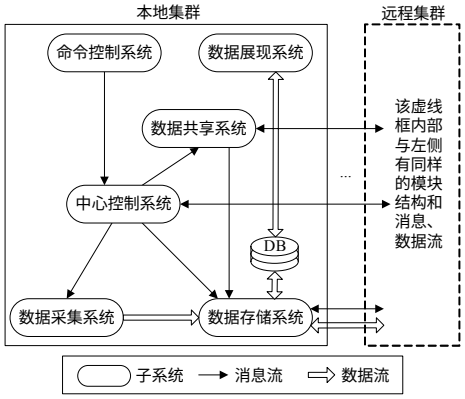

图 3　消息及数据流

4　运行实验分析

实验采用基于一维分解 Jacobi 迭代方程的 MPI 并行作业来进行。硬件环境为 IBM 刀片集群：SMP 节点，CPU 为双 Intel Xeon 3.2 GHz, 2 GB 内存，节点间用千兆以太网互连；节点操作系统为 Red Hat Enterprise Linux AS 3，多集群作业调度系统为经过改造后的 OpenPBS。为了便于实验部分的分析，本文只选取了 2 个处理单位(Processing Unit, PU)，每个 PU 代表并行作业中的一个任务执行线程。

图 4 显示了任务运行的集群和节点名称、2 个 PU 的信息和 CPU 利用率情况。上方曲线为 PU4633，下方为 PU4634。结合实时 I/O 吞吐率信息能够分析出：PU4633 执行的任务 CPU 占用率很高，属于典型的计算密集型；而 PU4634 处理的是 I/O 吞吐量较大的任务。

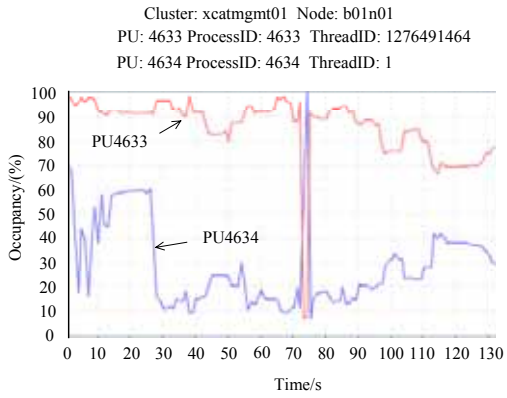

图 4　CPU 利用率实时曲线

采用模糊 C 均值聚类算法对性能数据进行聚类分析，结合本次聚类分析和数据库中的历史分析结果，自动设置一个合理的阈值，然后展现出经过阈值分析的结果，其性能指标统计如图 5 所示。其中，图 5(a)的阈值为 4 633.127 649 146 4，图 5(b)、图 5(c)为 4 634.1。

在图 5 中，并行作业中的任务被划分为计算密集型、I/O 密集型和等待密集型 3 类，工具把经过阈值分析后的 PU 分别归类，并以动态变化的图表形式将 PU 的性能信息显示出来。可以直观地看到，分析得到的结论与实时曲线所反应的特征是相符合的，性能监测分析工具能够较好地对多集群并行作业进行工作。表 1~表 3 分别列举了对超出 CPU, I/O, wait 利用率平均阈值的 PU 信息。

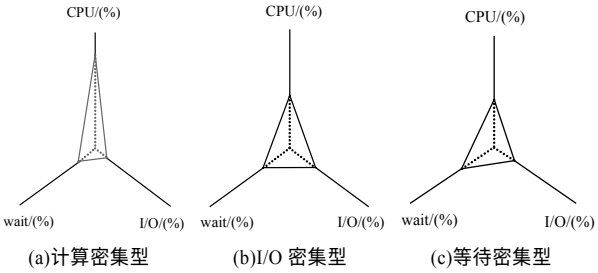

图 5　基于阈值的性能分析

表 1　超出 CPU 利用率平均阈值的 PU 信息

Cluster	Node	puID	Process	Thread	Value
xcatmgmt01	b01n01	4 633	4 633	1 276 491 464	0.830 557 591 612 902 8

表 2　超出 I/O 利用率平均阈值的 PU 信息

Cluster	Node	puID	Process	Thread	Value
xcatmgmt01	b01n01	4 634	4 634	1	0.318 056 089 232 257 9

表 3　超出 wait 利用率平均阈值的 PU 信息

Cluster	Node	puID	Process	Thread	Value
xcatmgmt01	b01n01	4 634	4 634	1	0.378 047 817 954 838 9

通过实验分析，也为后续研究工作提供了更多的思路，比如进行插桩对并行作业扰动方面的研究，不断丰富分析方法和展现形式等。这些研究将会在以后的工作中逐步开展。

5　结束语

本文结合多集群计算模式特点和动态性能监测分析方法展开研究，设计实现了一个并行作业性能监测分析工具。研究表明，针对多集群计算环境，采用分布式软件架构、模块功能完善的软件工具能够有效地对多集群并行作业进行性能监测分析，分析结果对提高并行作业的执行效率有很好的参考价值。

参考文献

[1] Roth P C, Miller B P. On-line Automated Performance Diagnosis on Thousands of Processes[C]//Proc. of ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming. New York, USA: [s. n.], 2006.

[2] Truong H L. Novel Techniques and Methods for Performance Measurement, Analysis and Monitoring of Cluster and Grid Applications[D]. Vienna: Vienna University of Technology, 2005.

[3] Stoica I, Morris R, Karger D. Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications[C]//Proc. of ACM SIGCOMM'01. San Diego, California, USA: [s. n.], 2001: 27-31.

[4] Project P. Dyninst Programmer's Guide[R]. Computer Science Department, University of Maryland, USA, 2006.

[5] 刘　建, 沈美明. Unix 进程文件系统及其在调试器设计中的应用[J]. 计算机工程, 2004, 30(4): 176-178.