文章编号:1671-6833(2005)02-0098-04

LINUX 集群系统并行应用程序监测技术的研究

王文义[1],梁青云[2],王若雨[3]

(1.中原工学院计算机系,河南 郑州 450007;2.郑州大学信息工程学院,河南 郑州 450052;3.河南电力职工大学网络中心,河南 郑州 450051)

摘 要:从大多普通用户的实际情况出发,在proc 文件系统基础上结合 MySQL 数据库和 GTK+ 技术,提出了一种简单实用的 LINUX 集群系统的并行计算监测工具的实现方法.该工具可以将运行中集群的节点状态实时地呈现给用户,其主要功能模块有用户交互、显示、控制、数据库服务端等 7 个模块,从而对进一步改进与提高并行应用程序的质量提供了科学依据.

关键词:集群系统;并行计算;MySQL 数据库;GTK+;proc 文件系统

中图分类号:TP 316.81 **文献标识码**:A

0 引言

计算机集群[1]系统,是指由网络形式互相连接在一起的计算机所组成的集合.其中的计算机可以是单机或多处理器系统(PC 或高性能工作站),它们可以有自己的存储器、I/O 设备和操作系统.目前,LINUX PC 集群系统可以说是最廉价的实用型并行处理计算机系统之一,由于它具有投资风险小、扩展性好、用户可继承的软件资源丰富且构造简单等特点,故已成为广大普通并行处理用户所乐于采用的方式.

过去,一般从事并行计算的集群系统用户,往往由于经费等各种原因,对各个计算节点的监测还多是采用手工方式.即在计算过程中,从主控机以命令行的方式检查计算状态.实际上,并行运算往往都需要系统作长时间的运行,因此在计算过程中不可避免地会出现节点故障或应用程序中断等现象,这样就会大大影响到工作效率.另外,用户通过用这种方式确定节点状态和进程状态,对于节点较少的小型应用系统来说还是可以的,但当系统规模扩展到几十个节点甚至上百个节点的时候,其难度就要增加许多,这也同样会造成工作效率的降低.

基于上述考虑,自行开发一种功能完善的 LINUX 集群系统并行应用监测软件是十分必要的.其突出优点是它与同等正版软件相比成本较低,而且可以将各个节点的状态实时的、准确的、清晰的呈现给用户,使用户能够及时发现问题、判断问题原因并解决问题,这样一方面可以尽量避免或减少因某个计算节点的瘫痪带来的损失,另一方面对进一步提高用户并行应用程序的质量也是大有裨益的.

1 监测系统现状及目标系统结构

LINUX PC 集群系统的监测功能应该具有监测对象准确、系统资源占用小和投资小的特点.目前国内外同类的监测系统大多数都是和硬件捆绑销售的,而且价格较高,如ISC(Intel Server Control)集群监测工具软件,仅适用于使用Intel 架构的带有集成管理功能主板的集群计算机,这对由普通计算机构成的集群来说是不适用的,而且对一般的集群用户来说也很难承受其高昂的费用.鉴于这些情况,本文作者提出了一种简单可靠的方法,以设计适合于LINUX PC 集群的监测系统,从而达到提高系统可用性的目的.

目标系统所提供的是一个统一的可视化图形管理界面,使用户能够在单一控制点上对集群各节点进行管理和配置.主界面内容应尽可能多地分类显示主要信息,以便使用户对机柜或某个节点的操作.为了构建一个功能强大、且易于使用和管理的并行计算应用监测系统,我们在系统框架上采用了较为可靠的Client/Server 模式,把被监

收稿日期:2005-01-28;**修订日期**:2005-03-20

基金项目:国家重点新产品计划资助项目(2002ED782017)

作者简介:王文义(1947-),男,河南省洛阳市人,中原工学院教授,主要从事并行计算机处理技术研究.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

测节点作为服务端(Server)，而把监测节点作为客户端(Client)．在服务端和客户端中分别含有用户交互模块、显示模块、控制模块、数据库服务端、接收指令模块、数据库客户端、数据采集等功能．监测系统的结构如图 1 所示．

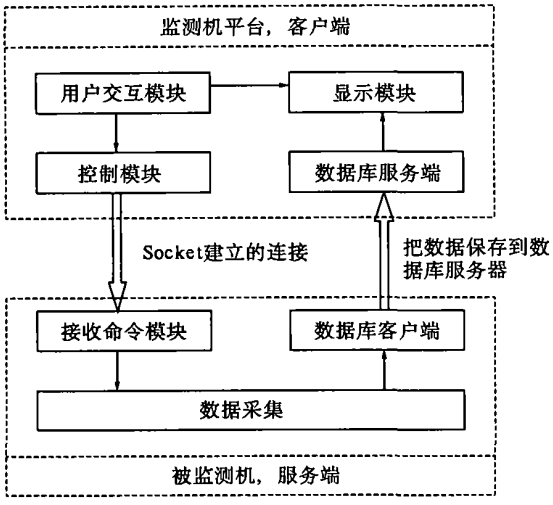

图 1　目标系统结构

Fig．1　Structure of object system

监测系统各功能简介：

(1) 用户交互模块．提供交互式的可视化人机图形界面来接受用户的指令用于管理整个集群．在模块中，系统可以接受用户的指令来完成显示和采集集群状态信息的功能．

(2) 显示模块．从数据库中提取各个节点的状态信息以在图形化界面中进行显示．

(3) 控制模块．完成监测节点依次向各个被监测节点发送指令的功能．被监测节点收到指令后，给予回应，建立Socket 连接．控制模块发送的指令是用来通知被监测检点来进行数据采集．

(4) 数据库服务端．并行计算中各节点在长时间运行时会产生大量的数据信息，这些数据信息可有选择的用来对并行程序进行分析、排错和优化．所以在并行计算过程中，节点信息是需要保存的，而数据库服务端则用于完成节点信息的保存功能．

(5) 接收指令模块．被监测节点的接收指令模块用来接收控制模块所发出的指令．当接收到指令后，被监测节点会按照要求采集当前节点机的状态，并把数据通讯并保存到数据库服务器端．

(6) 数据库客户端．完成节点数据信息向数据库服务端的保存工作．

(7) 数据采集．数据采集是整个监测系统的核心，用来采集节点机的各种信息和当前运行的进程状态．

通过以上各个模块的协同工作，可以把整个集群的状态动态的呈现给用户，从而完成集群的可视化管理．

2　系统的实现过程

在系统的实现上，前台界面采用C 语言和GTK$^+$设计，数据库系统采用MySQL，数据采集部分则在proc 文件系统上实现．节点机的控制通过向计算进程发送信号来完成．

2.1　监测界面的实现

由于监测系统要做到交互性和可视化，因此采用了在Xwindows 里流行的GTK$^+$界面函数库．GTK$^+$(GIMP TOOLKIT)，是一个跨平台的图形界面(GUI)开发工具，是目前LINUX 操作系统中常用的开发工具之一[2]．在程序设计中，我们只须重点考虑软件的应用功能，而软件的图形界面功能(菜单、工具条、按钮等)则可以通过调用GTK$^+$提供的库函数自动完成．

在并行程序运行期间，以一定时间间隔向指定文件中写入当前进程运行的进度数据，监测程序通过读取这些数据可以在图形界面中动态的以进度条方式显示计算进程执行的百分比例．而通过读取数据库服务器中的各个节点信息，就能够以图形方式把各个节点的当前运行状态显示给用户．

2.2　应用数据库的实现

由于大规模并行计算的时间一般都较长，所以监测程序需检测大量的信息．这些信息对并行应用程序的分析、排错和优化有着重要的作用，因此需要把这些数据保存到数据库中．在众多的数据库系统中我们选择了MySQL，因为它具有资源消耗低、速度快和支持网络存贮等功能；同时MySQL 还提供有用C 编程语言编写的客户机库，因此我们可以直接编写访问MySQL 的客户机程序[3]．在监测机上安装MySQL 的服务器端，这样被监测端就可以通过SQL 语句把数据保存到服务器端的数据库上．具体实现过程：

(1) 调用MySQL 的API 函数mysql_init()来获取连接处理程序．

(2) 调用mysql_real_connect()来建立客户端和服务器端的连接．

(3) 调用mysql_query("INSERT 或UPDATE 语句")完成向服务器端数据库保存数据．

(4) 使用mysql_close()来终止客户端和服务器端的连接．

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

2.3 系统节点状态的监测机理

2.3.1 网络状态判断

对节点状态的监测,我们首先要对网络的状态进行判断,简单的方法是使用ping 命令.但是ping 命令一次只能对一个节点(即一个ip 地址)进行监测.在对多节点进行监测时,就需要多次执行ping 程序,这会导致效率降低,况且在节点太多的状况下,这种方式是不能接受的.使用Socket建立连接可以较为圆满地解决这个问题,以一定时间间隔(缺省为25ms)向节点发送ICMP 请求报文;对已超时(超时时间的默认值为2500ms)但发送的请求报文数小于3个(默认值)的节点进行统计并发送下一个请求报文;对已发送了3个(默认值)请求报文仍然超时的节点就可以认为网络连接受阻.

2.3.2 并行计算过程的监测

proc 文件系统.对并行应用程序执行过程的监测是目标系统的关键.为了能够动态得到计算进程的实时信息,本文作者采用了LINUX 的proc文件系统.proc 文件系统是一个伪文件,它只存在内存中而勿需占用外部空间.proc 文件以文件系统的方式为访问系统内核数据的操作提供接口.用户和应用程序可以通过proc 文件得到系统和每个进程的信息,并可以改变内核的某些参数.由于系统的信息,如进程等是动态变化的,所以在用户或应用程序读取proc 文件时,proc 文件系统是从系统内核随时读出所需信息并提交的,因此通过proc 文件系统得到的是实时信息.系统中任何时刻正在运行的每个用户级进程在/proc 下都建有一个目录,目录的名称即进程号的十进制表示[4,5].如某个进程的进程号为2 345,则在/proc目录下面就会有一个名为2 345的目录,在这个目录里面存放着该进程的相关信息.proc 文件系统中的文件通常都是纯文本文件,使用fopen( )和scanf( )等函数就可以方便地得到进程信息.

节点机的监测.对于并行计算,监测软件需要监测的数据有系统负载状况、CPU 利用率、内存使用率和交换区使用状况等信息.而对于一个计算进程,需要监测的数据有进程对内存的使用情况、进程状态等信息.这些信息在proc 文件系统中可以通过读取相应的文件得到,分述如下.

(1) 负载状况.用fopen( )读取"/proc/loadavg"文件中的前3个数据,即代表系统在过去1分钟、5分钟和15分钟内的负载信息.如在/proc目录里loadavg 的文件内容形如:0.31 0.13 0.86 1/63 3258.

(2) CPU 利用率.读取"/proc/stat"文件,可以获取进程使用CPU 的统计信息.这些信息被分成用户时间(user)、有效用户时间(nice)、系统时间(sys)和空闲时间(idle).对该文件需要读取两遍,若用total 表示累计总时间,那么user +sys 就是我们想要知道的累计CPU 占用时间.每个变量后面的数字表示它是第几次读文件得到的,用如下方法可计算出CPU 占用率(由于两次读取的时间间隔比较短,我们可以近似地这样认为):

$$Total\_1 = user\_1 + nice\_1 + sys\_1 + idle\_1$$

$$Total\_2 = user\_2 + nice\_2 + sys\_2 + idle\_2$$

$$Rate = \{[(user\_2 + sys\_2) - (user\_1 + sys\_1)] / (total\_2 - total\_1)\} * 100$$

(3) 内存使用率.读取"/proc/meminfo"文件以获取系统内存的各种信息,通过空闲内存和总内存数即可得出内存利用率.

(4) 进程对内存的使用情况.读取"/proc/进程pid/statm"文件可以得到进程使用内存的大小、驻留大小和共享页面数等信息.

(5) 进程状态.读取"/proc/进程pid/status"文件可以得到进程的状态,R =正在运行,S =睡眠,D =不能中断的睡眠,T =被跟踪或停止,Z =僵进程,W =不驻留.

2.3.3 并行计算的过程控制

一个进程在执行的时候,因故可能会出现D状态(不能中断的睡眠)和Z 状态(僵进程).这两种状态说明当前的进程由于出错被挂起.当监测到并行计算进程出现了这些情况时,就要对计算进程进行相应的控制.本文作者采用的方法是向进程发送信号,也称软中断的方法,即在软件层面上对中断机制的一种模拟[6,7].监测软件通过向计算进程发送SIGKILL 信号来终止进程,同时向主计算进程所在的节点发送一个信息,通知主计算进程当前几号节点机的进程被终止,从而使主计算进程可以对计算进程进行从新分配并使计算继续进行.计算进程的动态分配只有在MPI -2 中才能实现.

通过上述方法,我们可以监测到计算节点的实时信息并能对其进行控制.由于直接从操作系统内核读取系统状态,因此保证了信息的准确高效.采用分布式结构,一台主控服务器可以监测多台服务器的状态.被监测服务器只负责数据采集,然后发送给主控服务器.尽量减少监测程序在被监测服务器上的系统开销.所有的逻辑判断、报警

(C)1994-2025 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

和数据存储等全部由主控服务器完成.

3　结束语

本文作者以 proc 文件系统为基础,结合 GTK$^+$ 和 MySQL,开发了一个基于 LINUX PC 集群的并行应用监测系统.该监测系统的成本较低、可用性好并且功能较强,比较适合于广大普通的集群系统用户.由于该监测系统主要是用于动态监测集群在执行并行应用程序时的运行状态,因此它可以及时地让用户发现问题并提供可靠的解决问题的依据,从而达到提高 LINUX 集群系统效率的目的.

参考文献:

[1] 陈国良.并行计算:结构、算法、编程[M].北京:高等教育出版社,2000.

[2] SMITH W.GTK$^+$ Reference Manual HOWTO[EB/OL].http://developer.gnome.org/doc/API/gtk,2001－01－13.

[3] PAUL Dubois.MYSQL 网络数据库指南[M].钟　鸣,译.北京:机械工业出版社,2000.

[4] KURT Wall.LINUX Programming Unleashed[M].王　勇,王一川,林花军,译.北京:清华大学出版社,2002.

[5] TREVOR Warren.Exploring /proc HOWTO[EB/OL].http://www.freeos.com,2000－12－12.

[6] 毛德操,胡希明.LINUX 内核源代码情景分析[M].浙江:浙江大学出版社,2002.

[7] 王文义,张　影.构建高性能集群计算机系统的关键技术[J].郑州工业大学学报,2001,22(1):6～9.

Study on Monitor System for Parallel Application on Linux Cluster

WANG Wen－yi[1], LIANG Qing－yun[2], WANG Ruo－yu[3]

(1.Department of Computer Science,Zhongyuan Institute of Technology, Zhengzhou 450007,China;2.School of Information Engineering,Zhengzhou University, Zhengzhou 450052,China;3.Network Center, Henan University of Electric Power and Works,Zhengzhou 450051,China)

Abstract: With the appearance fast development of the LINUX cluster system it is possible to develop the high performance computing tasks that can hardly be done before due to the limited conditions. The current monitor system software of parallel application program is so expensive in contrast to the cheap hardware resource of PC cluster system. This paper considers the factual situation that most users use the PC cluster system and bases on the proc file system and then utilize the MySQL database and GTK$^+$ technology to put forward a simple and applied method to implement the monitor tool of parallel computing. The tool can display truly and clearly every node's status on running cluster, so users can take further steps to improve the quality of parallel program.

Key words: cluster system; parallel computing; MySQL database; GTK$^+$; proc files system

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net