第34卷 第20期 计 算 机 工 程 2008年10月
Vol.34 No.20 Computer Engineering October 2008
·软件技术与数据库· 文章编号：1000—3428(2008)20—0034—03 文献标识码：A 中图分类号：TP309

基于SSI的远程集群管理系统

童 端，韩忠愿，苏杭丽
(南京财经大学信息工程学院，南京 210046)

摘 要：受集群系统结构的固有特性的影响，集群系统的管理问题日益突出。早期集群系统通过命令行方式进行管理，存在功能不完善、结构单一、可用性差、不支持远程管理等缺点。该文分析了集群管理软件的功能需求和相关技术，设计和实现了一套基于SSI的远程集群管理系统。该系统采用标准化模块设计方法，其功能可灵活组态，扩展性较好，并实现比较完整的单一系统映像，可提供简单、高效的管理功能。对系统进行了测试和评价，并提出该系统未来的研究方向。

关键词：集群管理；单一系统映像；远程管理

SSI-based Remote Cluster Management System

TONG Duan, HAN Zhong-yuan, SU Hang-li
(School of Information Engineering, Nanjing University of Finance & Economics, Nanjing 210046)

【Abstract】 With the popularization of cluster system, the management of cluster system becomes more important. In the early age, the cluster is managed by command line, therefore, some problems exist, such as imperfect functions, single structure, poor usability and nonsupport remote management etc. The conception of cluster management system and the technology related are analyzed in detail in this paper. Then, a SSI-based cluster management system is designed and implemented. The system is designed with the method of standardization module. It has flexible configuration functions and good scalability, and implements a relatively complete single system image. It also provides some kinds of simple and effective management functions. At last, test and evaluation are done for the system, and further researches are also introduced.

【Key words】 cluster management; single system image; remote management

1 概述

集群就是将一组相互独立的计算机通过高速的通信网络互连而组成的一个单一的计算机系统。由于集群具有性价比高和可扩展性好等优点，因此在并行处理领域中得到了广泛应用。

集群管理系统(资源管理软件)是对集群系统的各种软件资源和硬件资源进行管理的软件系统。由于集群系统固有的松耦合特性，集群中的节点在物理上是分散的，每个节点都有自己的一套系统，管理员想要协调一致地管理整个集群系统，就变得非常复杂和繁琐。用户使用集群系统时也会面临同样困难，这是影响集群系统推广和应用的一个重要因素。集群管理软件的目的就是简化集群系统的使用和管理工作，使集群系统简单易用、易管理。一般说来，集群管理软件应具备以下功能[1]：

(1)集群的安装与配置：组建集群系统是一项复杂的任务，要安装、配置必需的软件和硬件。因此，需要先进、易用的工具来完成这些工作。集群安装与配置软件可以根据用户的要求，自动安装所需软件，并完成集群的配置工作，大大减少了组建集群的工作量和复杂程度。

(2)软硬件的调度和分配：由于集群系统中多个节点要协同完成作业，使得作业管理和调度变得非常复杂，集群的调度软件就是要根据系统的资源状况，完成对作业的调度以及软硬件资源的分配。

(3)资源管理：资源管理就是分配系统的资源和监控系统资源的使用状态。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，甚至系统的事件和日志。

(4)监控与诊断：对持续运行的集群系统而言，当系统正常运行时，需要一些工具监控系统各部分的运行状态，如系统进程、CPU利用率和内存利用率等。

(5)分布式命令和文件：分布式命令和文件是指让命令和文件操作同时在整个集群节点或指定的一组节点上并行执行。分布式命令功能通常通过分布式的Shell来提供。分布式文件主要用于指集群中配置文件的同步。

2 单一系统映像SSI

单一系统映像(Single System Image, SSI)是一个虚像，是由硬件或软件创建的，使分散的资源集合起来作为一个统一的、更强大的资源使用[2-3]。SSI的概念可以用于应用程序、专用的子系统或整个集群。SSI对集群系统尤为重要。单一系统映像使得用户获得一个与登录点无关的全局视图，并为系统容错提供方便。基于集群的单一系统映像的设计目标集中在资源管理的透明性、性能的可扩展性以及系统的可用性等方面[2-4]。

(1)集群服务器的单一系统映像主要提供如下服务[5]：

1)单一入口点：物理上有多个节点处理用户的连接请求，但用户看到的是一台虚拟主机。

2)单一用户界面：用户能够通过一个单一的GUI来使用集群系统，就像用户在使用一台工作站一样。

基金项目：南京财经大学科研基金项目资助(B0614)
作者简介：童 端(1978－)，女，讲师、硕士，主研方向：分布式系统，供应链管理；韩忠愿，教授、博士；苏杭丽，副教授、博士
收稿日期：2007-10-29 **E-mail：**xtong_duan@sohu.com

3)单一进程空间：所有用户进程都有一个全局范围内唯一的 id 标识符，任何节点上的进程都可以在本地或其他节点上创建子进程。任何进程都可以与远程节点上的进程进行通信。集群系统应该支持全局的进程管理并屏蔽掉本地机与远程机的差异。

4)单一内存空间：用户拥有一个大的、集中的内存空间。物理上，这一空间是由多个分布的本地内存组成。

5)单一 I/O 空间：允许任何节点在本地或远程外设，包括磁盘上，执行 I/O 操作。通过这一服务，连接到节点机上的磁盘以及直接连接到网络的 RAIDS 和各种外设共同组成一个单一的 I/O 空间。

6)单一文件系统映像：用户看到的是一个巨大的单个文件系统映像，所有的目录和文件都在同一个根目录下。磁盘或其他文件系统设备对用户都是透明的。

7)单一虚拟网络：一个集群系统可以有多个不同网络。单一虚拟网络使得任何节点可以访问集群范围内的任何网络，即使这一网络并未与所有节点相连。

8)单一作业管理系统：通过全局作业调度器对提交到任何节点上的作业进行调度。作业的执行模式可以是批处理、交互或并行。

9)单一管理和控制点：利用单个 GUI 工具，通过一个窗口对所有节点进行监测、配置和控制。

10)检查点和进程迁移：检查点是一种软件机制，它定期地保存内存及磁盘中的进程状态和计算的中间结果。有了这些数据，发生错误时，系统可以卷回到检查点所记录的状态，从而进行错误恢复。进程迁移可以实现节点间的动态负载平衡，并为检查点机制提供支持。

在集群的SSI技术中主要的关键技术有单一入口点、单点管理和单一用户界面等[6]。

(2)SSI 的优点

1)在任一节点提供一个简单直观的、关于整个系统资源和状态的视图。

2)使操作者无须了解资源的物理位置，管理员可以在任一节点上管理整个集群。

3)为用户提供熟悉的接口和命令。

4)使最终用户不必知道程序在哪个节点上运行。

5)提供不依赖于位置的消息通信。

6)减少了对系统程序员的知识要求，简化了系统编程。

7)促进了标准的集群工具的开发。

3 集群管理系统的结构模型

3.1 系统的体系结构

常用的集群管理系统的体系结构模型有 4 种：集中式，分布式，分层式，分布式和分层式相结合的体系结构。在设计过程中，考虑到集群规模的变化较大，而且一般情况下，所有节点都在同一个局域网内部，本系统选择分层结构作为该管理系统的体系结构，如图 1 所示。

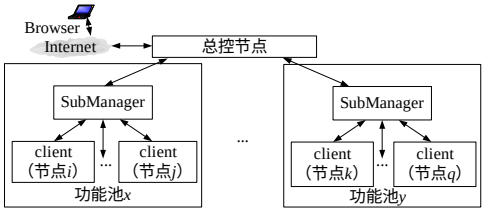

图 1 集群管理系统体系结构

3.2 系统的软件层次结构

基于上面的体系结构，整个系统可分为 5 层(如图 2 所示)，其中该管理系统包括最上面的 3 个层次：通信层，工具层，用户界面层。所有下层均为上层提供服务和支持。通信层用来提供集群系统中节点间的通信机制。工具层包括 3 个部分：系统安装工具，系统管理工具和系统监控工具，这一层是各项管理功能的具体实现。界面层提供管理工具的使用界面，这里使用的是基于 Web 的 B/S 结构。

图 2 集群管理系统分层结构

4 集群管理系统的模块化设计及具体实现

4.1 系统的模块化设计

按照系统中各部分所应具备的功能，该系统主要划分为如下几个模块：总控模块(Main Controller, MC)；子管理模块，也称节点池管理模块(Node Pool Management, NPM)；单节点管理模块(Node Management, NM)；底层执行模块 (Execute Module, EM)。模块结构如图 3 所示。

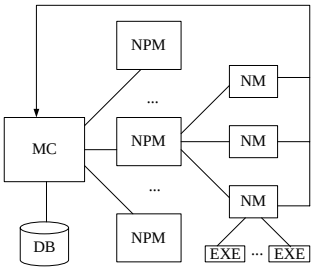

图 3 各模块之间的关系

(1)总控模块 MC：运行在控制台，对集群系统的信息进行组织、判断和管理。具体功能如下：

1)分析命令：判断要执行的是管理还是查询命令。

2)处理命令：根据客户端的请求，执行相应的命令，如果命令需要由 NM 来操作，那么把此命令转发给对应的 NPM。

3)定时查询每个 NPM 是否失效，若失效则报告给用户界面模块并将相关的故障信息写入数据库，并尝试启动一个新的 NPM 替代失效的 NPM。

4)端口数据的管理：设置管理系统的访问端口以及 MC, NPM 和 NM 的通信端口。

5)用户信息管理。

6)功能池信息数据以及节点信息数据的管理。

7)监听来自其他模块或相关软件的命令。

(2)节点池管理模块 NPM：运行在控制台或其他服务器节点上，是某一具体功能池的管理者，将 MC 的命令分解后转发给 NM，搜集各节点的性能信息并存储到控制台的数据库中。具体功能如下：

1)监听并分析 MC 的命令，以便及时处理其命令。

2)转发命令给一个或多个 NM。

3)向 MC 反馈结果信息。

4)定时查询 NM 的性能信息及其状态，并将性能信息存入数据库中。

(3)单节点管理模块 NM：运行在集群节点上，按照用户的命令对集群的节点进行管理，并监测节点的故障。具体功能如下：

1)监听并分析 NPM 的命令。

2)搜集本节点信息。

3)向 NPM 反馈结果信息。

4)监测故障并将故障报告给 MC。

(4)底层执行模块 EM：运行在集群节点上，执行具体的管理操作，向 NM 提供调用接口。具体功能如下：

1)供 NM 调用：NM 只是一个守护进程，当 NM 收到管理命令的时候，调用 EM 执行具体操作。

2)执行具体的管理操作：所有具体的管理和查询操作都通过该模块来完成。

3)向 NM 反馈结果信息：把命令的执行结果按照通信格式打包之后返回给 NM。

(5)信息数据库 DB：运行在前台节点上，存储系统的性能信息、功能池信息、节点信息、日志信息、故障信息、用户信息、设备信息等。

4.2 主要模块的实现

由于本文采用了模块化的体系结构，各模块之间的通信机制显得尤为重要，因此首先要设计一种用于各模块间通信的内部命令包通信格式。命令包格式由 2 个部分组成：参数部分和消息部分。参数部分由不同的项组成，每个项包括名称和值两部分并以“ = ”连接起来，各个项之间以“ ;”隔开。消息部分则是一个整体块，由参数部分的 nodenum 项和 msglen 项来区分各个节点所附的消息。

(1)控制模块

控制模块分为 2 个层次，上层是 MC，下层是 NPM。MC 在运行时是一个守护进程，对上层 UI 和其他应用程序来说是集群单一系统映像管理的一个接口。其工作主要有 4 个部分：初始化，监听消息，分析消息，处理消息。NPM 为 MC 的派出管理模块，同样是一个服务进程，针对某一具体功能池的管理，例如 Web 服务功能池模块、Email 功能池管理模块等。

这部分的主要技术难点包括：

1)NPM 失效时的发现和处理方法。

2)配置信息以及监控信息的存储和组织。

3)如何提高该模块的可扩展性和易用性。

NPM 和 NM 通过心跳向 MC 汇报自己的状态，对于 NPM 失效时的处理方法有 2 种：重启一个新的 NPM，或者直接由该总控节点直接把命令发往相应的 NM 模块。本系统采用重启一个新的 NPM 来进行信息的转发或反馈，以保证系统层次结构上的清晰。

控制模块关心的数据分为 2 类：各个节点池管理进程和节点管理进程的信息，包括类型、可以提供的服务、当前状态等，这些信息在每次管理调度的时候都需要查询，使用非常频繁，所以，需要常驻内存，采用树形结构进行管理和组织。而对于固定的配置信息和集群运行产生的历史数据，把它们存入数据库，方便以后的查询和分析。

所有的 NPM 模块和 NM 模块按照统一的通信接口规范进行设计，这样，随着集群规模的扩大和功能的增强，只要在数据库中增加配置信息，就可以不断有新的 NPM 加入进来，每一个 NPM 启动后都会向 MC 注册自己，由 MC 根据 NPM 类型和功能把它纳入自己的管理之下。而且，可以根据集群服务领域的不同，调整使用不同 NPM 模块，实现领域内的个性化需求。

(2)节点执行模块

所有的管理操作都需要依赖 EM 来执行。EM 向 NM 提供调用接口。EM 主要用 Perl 语言来实现。NM 通过堆栈的方式调用 EM。相互间的关系如图 4 所示。

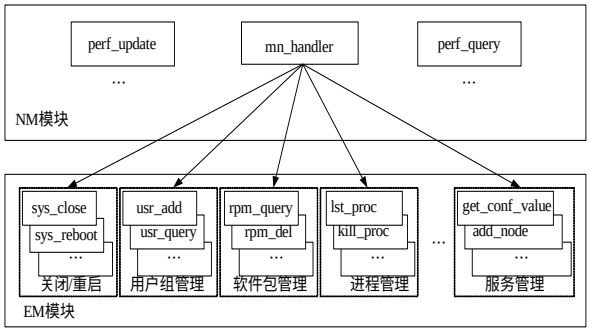

图 4 NM 和 EM 的调用关系

根据所提供功能，EM 可分为如下几类：关闭/重启服务器，用户组管理，软件包管理，进程管理和服务管理等。

5 测试及评价

5.1 测试环境

管理节点(总控节点)和 10 个集群内部节点均为浪潮英信 NF180(2 个 Xeon 2.4 GHz CPU、2 GB 内存、SCSI Ultra160，512 MB 内存，内置 2 个 1 000 Mb/s 网卡)，交换机为百兆交换机(3com 3c17700 12 端口)。操作系统选用的是 RedHat Linux 9.0 for ia32，数据库使用的是 MySQL。

5.2 测试结果与分析

进行系统性能测试时，管理单元和处理单元的 CPU 利用率和网络带宽的占有率如表 1 所示。

表 1 系统性能测试结果 (%)

单元	CPU 资源占用率	网络带宽占用率
管理单元	<0.5	<0.1
处理单元	<0.1	<0.1

测试结果表明，管理单元的 CPU 资源占用率是整个系统的瓶颈，当有 10 个节点规模时，管理单元 CPU 资源占用率小于 0.5%。按此推算，当系统扩展到 512 个节点规模时，管理单元的 CPU 占用率将小于 25%。

该系统的设计目标就是一个易用的单一映像的管理工具。它用 C++编写，同时采用了数据库和 Web 服务器等技术，使得集群系统的应用获得下面 3 个方面的提高：

(1)实现单一系统映像的管理功能，简化集群系统的管理难度，提高集群系统的可靠性和可用性。

(2)通过一个统一的图形用户界面，为使用者提供了一个更好的管理接口。同时，本文采用 B/S 的设计模式，方便进行远程管理。

(3)通过分层的体系结构设计，保证系统的高可扩展性。

6 结束语

本文首先提出了集群管理系统应该具备的几个基本功能，然后结合单一系统映像技术设计并实现了一个集群管理系统，使得用户可以从本地或远程通过不同的操作平台来使用该系统进行集群的管理，从而简化了用户的工作，提高了系统的安全性。基于 SSI 的远程集群管理系统是整个集群服务器系统的一个重要组成部分，随着网格技术的发展，本文也将在支持网格应用方面作一些尝试，使其适应新的网络环境。

(下转第 39 页)