

文章编号:1001-9081(2006)02-0485-03

## 虚拟化技术在基于自律计算的高可用性系统中的应用

刘文洁,李战怀

(西北工业大学 计算机学院,陕西 西安 710072)

(liuwenjie@co-think.com)

**摘 要:**针对服务器集群的管理复杂度高的问题,在对自律计算和大型可分区服务器硬件特征研究的基础上,提出了一种虚拟化技术,通过对硬件资源的虚拟化,实现了物理资源向逻辑资源的映射,从而统一管理所有物理资源。此外,在资源统一管理的基础上,对服务器集群的资源配置提出了完整的解决方案,实现了系统的自我管理。

**关键词:**自律计算;高可用性;可分区服务器;服务器集群;虚拟化

**中图分类号:** TP311 **文献标识码:** A

## Application of virtual mechanism in high availability system based on autonomic computing

LIU Wen-jie, LI Zhan-huai

(Department of Computer, Northwestern Polytechnical University, Xi'an Shaanxi 710072, China)

**Abstract:** To lower the management cost of server clusters, a virtual mechanism on the basis of studying autonomic computing and the hardware features of partitionable servers was proposed. By virtualization of the hardware resources, physical resources could be mapped to logical resources, so that they could be unified to manage. Moreover, on the basis of unification of resources, a series of measures were proposed to configure resources of servers cluster to achieve the self-management of system.

**Key words:** autonomic computing; high availability; partitionable sever; servers cluster; virtual mechanism

### 0 引言

企业用户在构建信息应用系统时一般采用的都是基于客户机/服务器的体系结构。这种计算体系允许用户根据实际需要逐步增加硬件系统,但缺乏必要的可用性和管理性。对于客户来说,管理水平还不高,服务管理的 TCO(瑞典专业雇员联盟就电磁场制定的更高标准)也随着系统规模成比例增长,这样对客户造成沉重的负担和很高的代价。同时,这种系统结构中存在很多问题,例如硬件故障后,SPOF(单点故障)状态时间长,需要花费很长的时间去校验集群系统的配置、确定硬件故障点、安装软件补丁或者寻找硬件/软件的变化,很难应付突然的工作负荷的加强等等。这些问题的解决需要更多的高水平的高可用性管理、更有效的硬件资源利用<sup>[1,2]</sup>。

IBM 在 2001 年 10 月提出自律计算的构想,指出自律计算系统具有四个主要特征:自我配置能力、自我修复能力、自我优化能力和自我保护能力<sup>[3,4]</sup>。自律计算环境的核心是使得 IT 系统迈向 RAS(高可靠性 Reliability、高可用性 Availability、高服务性 Serviceability)的目标。本文提出了一种虚拟化的技术,以大型可分区服务器为管理对象,通过屏蔽服务器硬件特征,将物理资源映射到逻辑资源上,实现了资源的统一管理和配置。并通过灵活的资源映射关系,实现了动态的资源重组,在此基础上,设计了基于自律计算的资源管理方案,从而实现系统的自我修复、自我配置、自我保护和自我

优化。该方案减少了复杂的异构环境下系统的管理成本,提高了系统的可用性,实现了系统的自我管理。

### 1 管理对象特征

本文采用虚拟化技术,实现了高可用性系统的管理软件,称作虚拟操作环境。

虚拟操作环境的开发,是基于某型号的可分区服务器。服务器为了实现高可用性,在硬件基础上采用了某些提高可用性的方法。

可分区服务器由一些机柜组成。每一个机柜为一个 32 路可分区的 SMP(对称多处理器)服务器,它拥有八个 Cell,每个 Cell 中有四个 CPU,通过对硬件进行分区,每一个分区可以运行一种操作系统(HP-UX/ Windows XP/ Linux),每个分区至少有一个 Cell,所以一个机柜能够有八个操作系统<sup>[5]</sup>。

通过分区技术,可将服务器内的 CPU、内存、I/O 等资源合理地进行分区和调配,不同分区内可以执行不同的操作系统或同一操作系统的不同版本,最大限度地挖掘了服务器的性能。

在虚拟操作环境中,我们管理的可分区服务器是由 16 个机柜组成,总共有 128 个分区,512 个 CPU。虚拟操作环境控制可分区服务器来管理共 128 个共享硬件资源的分区,通过有效地控制 512-way 硬件资源,虚拟操作环境可以把这个 512-way 的服务器当作简单的 128 个 4-way 系统节点来处理。

收稿日期:2005-08-22;修订日期:2005-11-06

作者简介:刘文洁(1976-),女,博士研究生,主要研究方向:软件理论、自律计算、高可用性系统;李战怀(1961-),男,教授,博士生导师,主要研究方向:数据库与知识库系统、存储区域网络、软件工程。

这是一种巩固服务器的新形式,我们把它命名为“服务器集群”。

该服务器有以下高可用特征:

- 1) 用 NUMA 可以达到 512 路实现系统的高可用性,节点通过高速的 Crossbar 相连;
- 2) 通过各个服务器和 Crossbar 的自由组合来支持节点间的高速通信;
- 3) 允许逻辑的配置来实现无单点故障(SpoF);
- 4) 使用分区概念来管理硬件资源,每一个分区可以运行一个 OS,可以对外提供一个服务,也称为一个 Server;
- 5) 提供了更高的可靠性和更好的 RAS 功能。

## 2 物理资源的虚拟化

虚拟化的基本思想是将物理资源映射为逻辑资源,用户所有的对物理资源的操作都间接的通过逻辑资源来操作,实际的物理资源对用户完全透明。物理资源向逻辑资源的映射是一对多的,针对数量固定,资源有限的情况下,对物理资源从逻辑上再次划分可以提高资源的利用率。本文从逻辑上将可分区服务器可分成若干个 Partition,每一个 Partition 都成为一个独立的服务器。因此通过逻辑虚拟化技术,一个高端服务器成为了若干个小服务器。所具有的特点:

- 1) 不同种类 OS(Operation System) 可以同时运行;
- 2) 各分区使用的硬件资源在系统运行中可以即时进行再分配;
- 3) 通过共享内存区域可以将分区连接成集群;
- 4) 集群之间通过自己专用的协议进行通信;
- 5) 引入自律计算技术实现系统的自我管理。

## 3 虚拟化技术的应用

以大型可分区服务器为管理对象,其硬件资源具有一定的复杂性,为了实现管理对象的自我管理,使其符合自律计算的特征,系统采用了虚拟技术,开发了一个易于管理的软件系统,称作虚拟操作环境,其主要目的是在对用户隐藏其复杂性的前提下,实现服务器的各种各样的功能。让系统实行一定程度的自我管理,从而降低服务器管理的复杂性,提高服务器的使用效率。

### 3.1 系统的体系结构

我们提出的虚拟技术的主要目标是为了屏蔽硬件,提高可用性。对物理的硬件资源进行屏蔽,用户只能通过逻辑资源进行操作,这样减少了出错的可能性。通过实现自我修复,自我调整,自我保护的技能,逻辑的资源要提供无停止性,服务级别保证,易管理和安全保证的功能。

为了解决系统的可用性和管理问题,首先需要解决:自我保护、自我调整、自我修复。虚拟操作环境提供了对上述问题的解决方案,实现了系统的高可用性和易管理性。

采用虚拟化技术实现的高可用性系统的体系结构如图 1 所示。

虚拟操作环境是为了自动灵活的控制可提供分区服务的计算机的分区构成变化而设计的软件。它主要是基于分区技术。

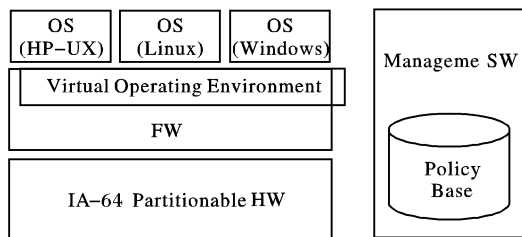


图1 虚拟操作环境体系结构图

由图 1 中可以看出,在虚拟操作环境系统中可以同时运行多个不同的操作系统,这些操作系统通过虚拟操作环境对可分区的硬件进行管理。硬件的配置管理完全由软件系统来完成。虚拟操作环境系统实际上是位于操作系统之上的配置控制管理模块。而该系统采用和实施的所有管理和配置控制都是基于策略库来完成的。这样使得系统的各项实施动作都是基于策略库中所定义的策略来完成的,从而保证了系统运行的科学性,提高了整个系统资源的利用率。

### 3.2 系统模块构成

为了在各个真实的服务器上执行指令(包括负载的收集等)需要在各个分区上有一个代理。而且各个真实服务器之间的通讯用一个集群通信驱动器实现。系统控制的接口是通过一个固件来实现。当然,对管理员来说,控制台是它的武器。因此,虚拟操作环境要由以下的模块构成:虚拟操作环境代理;机柜内的集群通信驱动器;虚拟操作环境控制固件;虚拟操作环境管理控制台。

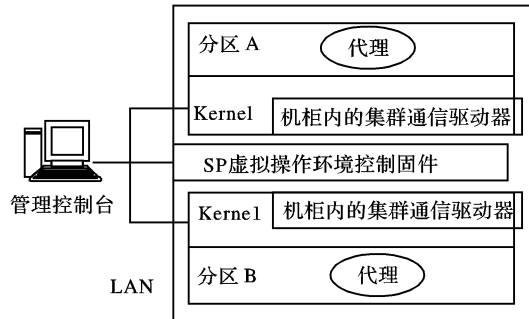


图2 虚拟操作环境模块构成图

SP 是内置的小型处理器,为了强化服务型计算机的 RAS 等功能。以 intel 为代表的各个代理商共同决定了 IA-32 和 IA-64 两种体系结构计算机的 SP 和本体的接口,以及和外部的管理者的接口,这些被定义为 IPMI,智能平台管理接口。

拥有独立的 OS 和外部接口,通过独立的 LAN 和外部的管理者(console)进行通讯。

虚拟操作环境代理存在于管理对象的可分区服务的分区中,实际上是这个分区中运行的操作系统中的一个后台守护进程。

机柜内通信驱动器存在于管理对象可分区服务的分区中,是这个分区中运行的操作系统内核中的驱动器。

虚拟操作环境控制 FW 是 SP 的 FW 的一部分,它作为管理对象,在可分区服务器的 SP 中操作,也是对系统处理器进行管理操作的固件的一部分。尽管它主要是虚拟操作环境管理控制台的一个接口,但也是整个固件的一部分,提供一般的系统管理功能。

虚拟操作环境管理控制台就是所谓的管理对象系统,由

独立的计算机构成,通过 network(通常是 LAN)连接。管理对象系统通过网络连接各个分区和 SP,这种构成,使虚拟操作环境管理控制台可以直接与各个分区中的虚拟操作环境代理通信,也可以直接与 SP 通信。

### 3.3 系统资源配置方案

为了使系统具有自我管理能力,达到高可用性目标,在实际的设计中,必须结合系统的硬件基础,从软件的层次上来进行设计。虚拟操作环境提供了一系列的资源配置方案,来实现系统的自律管理。

#### 3.3.1 动态的资源重组

一般的,在大型的管理系统中,系统的资源的需求是随着时间和系统的负载的变化而不断变化的,所以系统的资源的分配也应该是随着系统的状态变化而动态的进行调整,这样才能保证系统资源的最大利用率,提高系统的可用性<sup>[6,7]</sup>。

在虚拟操作环境系统中,将所有可能影响系统资源的因素存储在知识库中。用户可以定义自己的因素,这都称为策略。系统主要按照知识库中的策略来启动系统的资源重新分配。

#### 3.3.2 根据系统负荷进行资源重组

虚拟操作环境系统中,要能够及时,准确的收集负荷情报,并且根据负荷的大小,按照知识库中的策略要求,来进行资源的重新分配。

例如,当系统监测到某一个分区的当前负荷已经达到或超过分区的阈值上限的时候,这时,系统会自动的给该分区再进行资源的分配,一般是添加一个 CELL。这样增加了这个分区的处理能力,避免了因为资源不够而导致分区不能完成任务。同样的,如果在某一时间内,分区的负荷非常小,已经小于该分区的阈值下限,而系统中当前的预备资源已经分配完,系统又存在负荷超过阈值上限的分区,系统则剥夺负荷小的分区的资源分配给负荷大的分区,以实现系统的负载平衡。图2描述了这一情况。

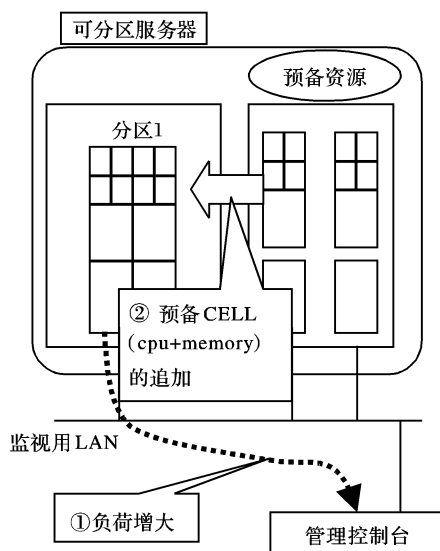


图3 负荷变更时资源重组示意图

当系统的资源处于紧张状态时,比如,预备资源已经用完,处于高负荷状态的分区又有很多,就整个系统而言,资源是不充足的,不可能对所有的高负荷状态的分区进行资源添

加,这时在系统进行负荷平衡时,按照分区上所运行的业务的优先级进行资源的剥夺与重新分配。

#### 3.3.3 周期性的资源重组

虚拟操作环境允许用户根据自己的业务需求来设定系统数据库中的时间表,系统根据数据库中所定义的周期来重新配置系统的资源。

例如,对一般的用户来说,系统在白天的任务主要是进行在线业务的处理,而数据备份的任务少一些,这个时候,相应的在线业务服务器的资源应该分配的多一些,数据备份服务器的资源要少一些。而在晚上,业务处理的业务少一些,系统主要进行白天业务的数据备份,那么,数据备份服务器占用的资源就要多一些,业务服务器的资源要少一些。

#### 3.3.4 预备资源的自动定期检测

在系统进行动态的资源分配的时候,预备资源扮演着很重要的角色,可以说,动态的资源分配是基于预备资源可用的基础之上的。那么,预备资源的可用与否,直接影响着系统能否实现动态的资源分配。

在虚拟操作环境系统中,系统对预备资源进行自动的,周期性的检测,以保证预备资源的高可用性。而且,周期性的用预备资源替换正在使用的资源,减少正在使用资源的硬件故障的发生几率,从而保证了系统的高可用性。

## 4 结语

本文采用了虚拟化技术,以大型可分区服务器为管理对象,通过屏蔽硬件特征,实现了物理资源向逻辑资源的映射,从软件的层次对资源进行管理,减少出错的可能性,从而提高系统的可用性。通过设计各种资源配置方案,可以实现系统的自我管理。该系统已经投入实际应用中,性能稳定,能够有效减少用户的干预。但是由于业务需求和管理对象都有可能发生变化,所以目前设计好的系统还存在很多需要改进的地方,有待于进一步研究和开发,这将成为今后的研究课题。

### 参考文献:

- [1] BUYYA R. 高性能集群计算: 结构与系统[M]. 北京: 电子工业出版社, 2001. 15-17.
- [2] 方华锋. 浅谈集群技术的应用[J]. 计算机世界, 2000.
- [3] HOM P. Autonomic Computing: IBM's Perspective on the State of Information Technology. IBM Corporation[EB/OL]. <http://www.research.ibm.com/autonomic> Oct, 2001.
- [4] MAINSAH E. Autonomic computing: the next area of computing[J]. Electronics & Communication Engineering Journal. 2002; 1-3.
- [5] HP Company. HP System Partitions Guide Administration for nPartitions Twelfth Edition [EB/OL]. <http://docs.hp.com/en/5991-1237/ch02s03.html#aes-npar-338>. 2005.
- [6] 傅强, 郑伟民. 一种适用于机群系统的任务动态调度方法[J]. 软件学报, 1999, 10(1): 19-23.
- [7] MARC H. WILLEBEEK - LEMAIR, Strategies for Dynamic Load Balancing on Highly Parallel Computers[J]. IEEE Transactions on parallel and distributed systems, 1993, 14(9): 979-993.