第 33 卷 第 5 期 计 算 机 工 程 2007 年 3 月
Vol.33 No.5 Computer Engineering March 2007
·网络与通信· 文章编号：1000—3428(2007)05—0079—03 文献标识码：A 中图分类号：TP338.4

一种提高高性能服务器资源利用率的新方法

李景山[1,2]，庄文君[2]，周立柱[1]
(1. 清华大学计算机科学与技术系，北京 100080；2. 浪潮集团，济南 250000)

摘 要：目前高性能服务器能够支持多种应用，但各种应用在同一段时间内对资源需求是不均衡的，致使高性能服务器的资源利用率较低。该文所提出的动态部署系统是一种提高基于 InfiniBand 和 SAN 的高性能服务器资源利用的新方法。该动态部署系统通过构建虚拟服务器，改变服务器结点的计算特性，在各种应用中移动计算节点，从而合理分配服务器中闲置资源，提高资源利用率和应用的性能。对动态部署系统的基本概念、设计、功能实现、性能测试及其理论证明等进行了研究。

关键词：高性能服务器；动态部署；虚拟服务器

Novel Approach to Improve Resource Usage in HPS

LI Jingshan[1,2], ZHUANG Wenjun[2], ZHOU Lizhu[1]
(1. Department of Computer Science and Technology, Tsinghua University, Beijing 100080; 2. Langchao Group, Jinan 250000)

【Abstract】 Now HPS(HPC) can support many application, but the resource usage of HPS is very low. Dynamic deployment system——a novel approach to improve resource usage in HPS is presented based on InfiniBand and SAN. Administrator can change the single computer's or whole HPS's computing feature in fly and move computing nodes in HPS easily with the dynamic deployment system. In this new type HPS, it can improve the usage of the whole HPS and improve the performance of partly HPS (smaller HPS) by adding the idle resource to busy application. The basic conception, design and implementation, and the theories of dynamic deployment system are discussed.

【Key words】 High performance server(HPS); Dynamic deployment; Virtual server

随着功能强、价格低的工作站和高速网络设备出现，机群逐渐成为高性能服务器的主流[1]。机群可以提供诸如网络服务、数据处理、电子邮件处理等多种应用。目前，如果想改变结点所支持的应用，必须重新安装相应结点的操作系统或者相应的支撑软件，这样改变结点应用特性的方式是非常费时费力的。本文提出了一种把计算资源和存储资源分离再动态地绑定计算与存储两种资源构建虚拟服务器的方法，使得系统能够容易地改变计算结点的计算特性，灵活提供多种应用。该系统能够将高性能计算机中的闲置结点集中起来；将闲置服务器与存储资源绑定，使之成为按需而变的虚拟服务器，快速改变计算结点的阶段特性，支持多种应用；从应用中增加或删除计算结点，提高应用性能，保证应用的服务质量。

1 动态部署系统的基本原理

1.1 计算与存储共享

随着操作系统网络化的发展，使得单个计算机的体系结构发生了很微妙的变化。逻辑上可以把计算机分成 CPU、NIC、DISK 3 部分，这样能够让这 3 部分在机群形式的高性能服务器中进行资源共享，把 CPU 和 DISK 动态绑定，使之成为虚拟的服务器，迅速改变计算结点的计算特性，为不同的应用提供支持。更进一步，可以将高性能服务器分解成许多小的高性能的计算机，根据不同的需求增加或删除虚拟服务，使得这些小的计算机的性能按需要动态变化，提高单个应用的性能和服务器整体的利用率。

1.2 高性能服务器体系结构的新观点

在高性能服务器中，通过分离计算和存储资源并且动态地将两种资源绑定，可以动态地构建计算机，迅速地改变其计算特征。基于这种思想，能够让高性能服务器提供多种应用，这些应用把一台巨大的服务器分割成许多小的计算机，这些小的计算机可以通过增加或者减少结点的数量动态地改变其计算性能和规模。动态部署系统通过把空闲资源加到资源紧张的应用中，能够提高整个高性能服务器的资源利用率和局部高性能服务器的性能。如图 1 所示，高性能服务器被应用分割成多台服务器，多台服务器的性能可以通过动态部署系统动态调整，按照用户的实际需要，各个局部的服务器性能可以提高也可以降低，计算资源和存储资源可以方便地在各个服务器之间移动，使资源得到最大限度的使用，提高了服务器的整体利用率。

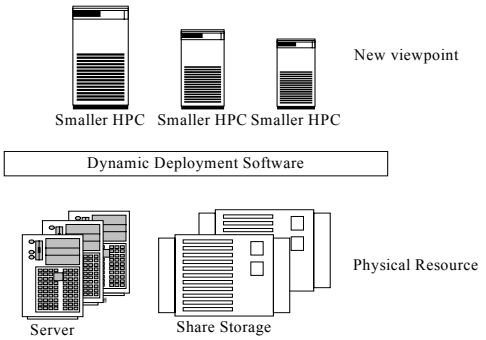

图 1 高性能服务器体系结构新观点

基金项目：国家“ 863 ”计划基金资助项目“ 新型网络服务器（一）”(2004AA111110)
作者简介：李景山(1973－)，男，博士后，主研方向：高性能服务器系统软件，网格计算，普适计算；庄文君、周立柱，教授、博导
收稿日期：2006-04-30 **E-mail：**lijsh@langchao.com.cn

2 设计与实现

2.1 天梭动态部署系统的 3 层结构

天梭动态部署系统软件由 Java 语言编写，由 JSP、Java Bean 和 Taglib 完成 Web 应用程序，提供基于浏览器的访问界面。如图 2 所该软件分成 3 个层次：Web 界面层，业务逻辑(Business)层，数据处理(Data)层。采用流行的 BS 体系结构完成了计算资源管理、存储资源管理、虚拟服务器构建、应用管理功能，实现了基于 Infiniband 网络和 SAN 的动态部署系统。

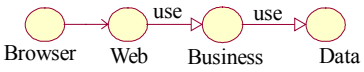

图 2 动态部署 3 层结构

2.2 计算资源管理

动态部署系统发现系统中的物理服务器，把这些物理服务器作为共享的计算资源。计算资源的管理包括物理服务器的发现、删除、修改等功能。通过计算资源的管理，动态部署系统完成了对空闲计算资源的管理。

2.3 存储资源管理

在天梭超级服务器中，每个节点自身都带有局部硬盘，这些硬盘被每个结点独占，它们装有操作系统和相关软件。在所有计算节点之间有一些共享的存储资源，如 NFS 服务器或 SAN。NFS 服务器为成本低、性能低的存储器，SAN 为成本高、性能高的存储器，它们都属于共享存储系统，在以后的章节中如无特殊说明，无须区分。通过在这些共享存储上创建、删除、复制操作系统和应用程序的映象，实现了对存储资源的管理。

2.4 构建虚拟服务器

构建虚拟服务器的关键问题是如何在 NFS 服务器或 SAN 上引导和运行操作系统。目前有一些运行操作的解决方案，在网络文件系统服务器上建立虚拟服务器是比较容易的，可以用在 NFS 服务器上建立无盘工作站的方法构建基于 NFS 类型服务器的方案。

也有从 SAN 引导的方法，但这些方法只适用于结点直接与 FC-SAN 连接的情况。通常的服务器或者刀片服务器如果要从远程的 SAN 引导时，要求连接存储的 HBA(Host Bus Adapter)卡或者 HCA 具备从所连接的设备引导操作系统的功能特性，这对 HBA 卡、HCA 卡和相应的交换机提出了更高的要求。通过对多家 HCA 卡和交换机的测试发现，在天梭计算结点通过 HCA 卡并不能实现操作系统的网络引导。这也说明我们研制的系统不应该依赖于特定厂商的 HCA 卡、HBA 卡和相应的交换机。利用 Linux OS 的运行特征，即引导操作系统和运行操作系统分为 2 步完成。主要过程为，虚拟服务器先从带有 PXE 支持的网卡下载 NFS 服务器操作系统核心及其 initrd 程序，利用 initrd 中的程序加载 HCA 驱动，虚拟服务器识别出 SAN 的存储信息，虚拟服务器把根文件系统等切换到 SAN 上，开始在 SAN 上运行操作系统、服务及其应用等程序。通过以上的步骤，本文实现了构建虚拟服务器与 HCA 网络引导无关技术，同时兼顾了虚拟服务器在高速 infiniband 网络上运行的优点。

2.5 应用管理

天梭动态部署操作系统允许用户在应用程序上增删虚拟服务器，应用程序可以在机群上运行，并能提供给用户界面或 API 来增删计算节点。我们选取了二类应用软件作为天梭动态部署系统软件的典型应用程序：(1)Oracle 10g RAC 商用软件，它可以在机群上运行，并可以在不用重启程序的情况下让用户增删计算节点；(2)在 Linux 平台上自主开发的机群单一 IP 系统软件－虚拟网络服务系统，这一软件的功能体现在它利用网络聚集协议能提供多个接入设备，允许多达 16 个接口设备并行工作，大大增强的网络访问量，天梭虚拟网络服务器提供 API 程序来增删网络接入结点。本文使用以上两种应用软件完成增删除虚拟服务器的操作，从而提高或者降低应用性能，提高资源的利用率。

2.6 天梭动态部署系统实例

天梭动态部署操作系统软件在天梭 30000 超级计算机上开发完成。这一软件能够查找闲置服务器并绑定相应资源，构建虚拟服务器，能自动或手动地将虚拟服务器加到相应的应用程序上。这一操作系统软件为用户提供了一种关于如何使用高性能计算机的新观念。如图 3 所示，用户可以很容易地按照需求将服务器从计算机 1 移到计算机 2 中去，或将其它闲置的服务器也如此移动。在下面性能评测部分会看到，动态部署操作系统可以加强整个高性能计算机的利用率和局部计算机性能。

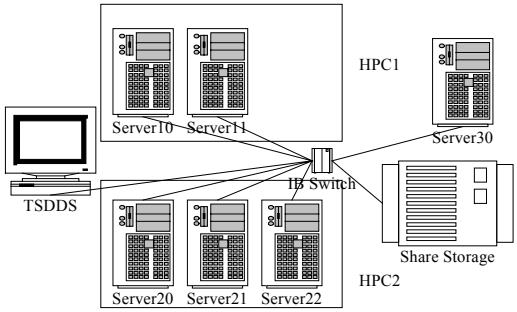

图 3 动态部署系统实例

3 动态部署中的 Amdahl 定律

本部分利用 Amdahl 定律来研究提高服务器整体利用率和局部应用计算性能的理论证明。

3.1 Amdahl 定律

Amdahl定律[2]是建立在固定工作负载基础之上的，多个处理结点并行计算作用于工作负载的可并行部分，用于缩短解决固定工作负载问题计算时间的模型。模型如下：

假设：总的工作负载

$$W = W_s + W_p \quad (1)$$

目标：使得计算时间最短，则加速比为

$$S_n <= W_s + W_p/(W_s + W_p/n) \le W_s + W_p/W_s = 1/f \quad (2)$$

公式中参数的含义：为计算结点的数量，W_s为工作的顺序计算部分；W_p为工作的可并行计算部分；S_n为加速比；f为顺序工作负载的比例。

3.2 Amdahl 定律在天梭动态部署操作系统中的含义

在天梭动态部署操作系统中，高性能计算在应用层面能够分成许多小的、局部的计算机系统，阿达姆定律只适应于小的、局部计算机中系统。通过借助 Amdahl 定律的分析手段，研究动态部署系统是如何提高整个高性能计算机的资源使用率和加强局部计算机的计算能力的。

假设将高性能计算机分为HPC_1和HPC_2两部分，分别有n个节点和m个节点($n >> m$)；每部分的工作量为W；为了论述方便，设$W_s = 0$(如果$W_s != 0$，得出的结论同样正确)，S_n为整个计算机的加速比，S_{n1}为HPC_1的加速比，S_{n2}为HPC_2的加速比。这样，使用动态部署系统之前：

$$S_{n1} = W/(W/n) = n \quad (3)$$

$$S_{n2} = W(W/m) = m \quad (4)$$

$$S_n = MIN(S_{n1}, S_{n2}) = m\ (n>>m) \quad (5)$$

S_n取S_{n1}，S_{n2}中较小的数值，当性能低的HPC_2仍在工作时，性能高的HPC_1已经完成工作，空闲下来。

使用天梭动态部署操作系统后，系统会根据应用运行的状况把 HPC1 的节点移到 HPC2 上，使每台机器有(m+n)/2 个节点。

$$S'_{n1} = (m+n)/2 \quad (6)$$

$$S'_{n2} = (m+n)/2 \quad (7)$$

$$S'_{n} = (m+n)/2 \quad (8)$$

3.3 结论

如式(4)和式(7)所示，局部机器的运算能力增强了，提高了局部计算机的加速比。式(5)和式(8)所示，整个机器的资源利用率提高了，提高了整体计算机的加速比。

通过天梭动态部署操作系统避免高性能计算机资源利用的不平衡性，实现计算机的资源在许多应用程序中共享。通过对 Amdahl 定律的计算资源动态调整情况下的重新诠释，在理论上证明了动态部署系统可以提高局部应用的性能和计算机整体的资源利用率。天梭动态部署操作系统为用户提供了基础的计算设施以实现计算机资源共享。

4 性能测试

本部分通过试验数据验证了虚拟服务器的性能以及在应用中添加计算节点后应用提高的性能。

测试环境如下：计算结点采用浪潮 NF420 服务器，该服务器具有两路 2.4GHz 至强处理器，2GB 内存，1 个转速 10 000RPM 的 36GB SCCI 硬盘。结点之间通过 1Gbps 以太网和 10Gbps Infiniband 网络互联，通过 Topspin 360 Infiniband 交换机和 EMC Clariion CX300 网络存储系统连接。结点上运行的操作系统是 Redhat AS 3.0，内核版本为 2.4.21，文件系统采用 ext3 文件系统。

4.1 虚拟服务器和物理服务器性能对比

本文利用 Benchmark Factory 中的 FileBench 基准测试程序测试了基于 NFS 类型和基于 SAN 类型的虚拟服务器相对与物理服务器的性能，本文采用的读写模式是从 1GB 大小的文件中读取块大小为 64KB 数据，写入块大小为 2KB 的数据到 1MB 大小的文件中。

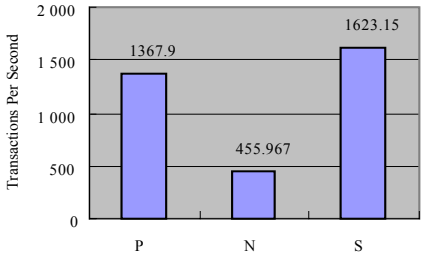

图 4 物理服务器与虚拟服务器性能对比

图 4 中，P 为物理服务器，N 为 NFS 类型虚拟服务器，S 为 SAN 类型虚拟服务器。通过图 4 的测试结果可以得出以下结论：NFS 类型的虚拟服务器的性能比物理服务器的性能低，而 SAN 类型的虚拟服务器的性能是物理服务器性能的 119%，超过了物理服务器的性能。通过动态部署系统部署的 SAN 类型的虚拟服务器完全可以替代物理服务器完成相应的计算任务。NFS 类型的虚拟服务器在 FileBench 测试中性能较低，是因为它使用了远程 NFS 服务器上的 NFS 文件系统，表现出文件读、写性能较差，但是可以利用该类型的虚拟服务器具有独立的处理器和内存的特点，让其承担计算密集型任务。

4.2 应用性能对比

本文利用 Benchmark Factory 中的 AS3AP 基础测试程序对并行数据据 Oracle 10g RAC 进行了测试。在并行数据库应用中添加一台虚拟服务器后，测试了该数据库应用性能提高的结果。

如图 5 所示，在 RAC 中只有一个结点时，在客户数量不断增长的过程中，用于结点的利用率达到饱和，TPS(每秒事务处理数)增加到 2 100 后，不能继续增加。利用动态部署系统后，系统可以快速地构建一台虚拟服务器，加入到 RAC 中去，图 5 中表现了在客户量达到 300 时，加入了一台虚拟服务器结点，RAC 的性能立即得到了提升，提高的幅度从 2 100TPS 到 3 400TPS，大约提高了 62%。

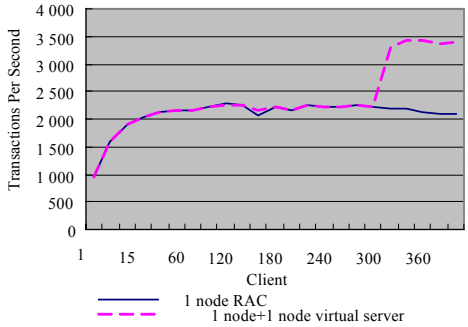

图 5 RAC 性能提高

通过以上两项测试，可以证明动态部署系统能够快速构建虚拟服务器，改变计算结点的计算特性，向资源使用紧张的应用中添加计算资源，提高该应用的性能，同时也提高了高性能服务器整体的利用率。

5 相关工作

5.1 IBM 的动态逻辑分区

动态逻辑分区[3](DLPAR)是AIX 5.2 的一大特色，通过引入动态逻辑分区的方法来增强逻辑分区的性能，这种方法能使一个分区的资源转移到另一个分区，而不需要重启系统程序，也不会影响其他分区的运行。动态逻辑分区能够把单个服务器分成许多可以独立运行的逻辑服务器。动态逻辑分区不提供构建拥有若干个虚拟服务器的高性能计算机的功能。

5.2 Topspin Vframe

VFrame[4]是由Topspin 公司开发的软件，VFrame基于InfiniBand技术，通过控制InfiniBand交换机，灵活地把物理服务器和存储映象绑定到一起，形成虚拟服务器。VFrame软件依赖于该公司独有的InfiniBand交换机、HCA及其一些自有的技术，不能在其他公司的InfiniBand网络上构建虚拟服务器， VFrame也不能提供应用管理功能，没有支持的应用实例。

6 结论

本文研究了天梭动态部署操作系统的原理、设计、实现、理论证明及其性能测试。这一系统能可以让用户不必重新安装新的操作系统和应用软件，就可以瞬间改变高性能服务器计算结点的计算特性，提供应用需要的功能，提高局部应用的性能和高性能计算机整体资源利用率，实现了高性能服务器功能特性根据应用需要快速按需而变的要求。

（下转第 90 页）