



一个集群文件系统的虚拟化支持平台研究与实现^{*}

邬延辉,陈 征,龚松春

(宁波大学信息科学与工程学院 宁波 315211)

摘要

在分布式数据管理中,服务器的虚拟化是一项重要工作,将网络中的多个服务器进行虚拟集成,可以快速完成一些大型的计算任务,并且减少硬件花费,减少空间和能量消耗。多数服务器虚拟化产品具有动态迁移的特点,动态迁移技术可以使虚拟服务器在运行时从一台机器转移到另一台机器,避免因硬件维护所产生的暂时或长时间服务中断。动态迁移需要共享存储器,其性能随虚拟服务器数量增加,成为整个系统的瓶颈。提出和实现了一个基于 Lustre 集群文件系统的、建立在开源虚拟机解决方案 KVM 之上的服务器虚拟环境。实验结果证实,虚拟环境提升了虚拟服务器的输入输出性能,同时使得创建 KVM 用户的耗时间明显缩短。

关键词 集群文件系统;服务器虚拟化;KVM 虚拟机

1 引言

服务器的虚拟化给分布式数据管理带来了巨大的、革命性的改变。通过虚拟化手段集成服务器能够有效减少空间和动力的消耗,提高信息服务的连续性,也能够简化硬件设备的管理工作。从信息服务的连续性角度分析,服务器虚拟化平台提供了一个有效的解决方案。当前 90% 的服务器关机操作是计划内的关机,目的是进行服务器设备维护。通过使用虚拟化平台的动态迁移技术,计划内的服务器关机次数能够显著减少,动态负载平衡使得服务器虚拟化更加强有力,将一些低负载的服务器进行合并,可以使用更少的服务器来工作,从而节省电能的消耗。

近年来,服务器虚拟化获得快速发展。在硬件方面,服务器 CPU 的一项基本功能就是支持虚拟化技术^[1],存储设备和网络设备同样提供服务器虚拟化的支持。在软件方面,服务器虚拟化软件的解决方案已经成熟。目前被广泛

使用的企业服务器虚拟化产品包括 VMware Server^[2]、Citrix XenServer^[3]、Microsoft Hyper-V Server^[4]以及开放源代码的 KVM(kernel-based virtual machine)^[5-7]。VMware 是服务器虚拟化领域的领跑者,它的虚拟服务器管理软件提供了许多有用的功能,但同时它的价格非常高。Citrix 的 XenServer 系统是 VMware 强有力的竞争者,其市场占有率成长速率很快。微软的 Hyper-V 也在快速发展中,Hyper-V 虚拟机管理程序被包含在了 Windows Server 2008 操作系统中。

与这些昂贵的商业软件相比,Linux 下的开源服务器虚拟化软件 XEN hypervisor^[8]和 KVM 是免费使用的。出现于 2007 年的 KVM 是虚拟化市场的后来者,依靠 CPU 对虚拟化的支持和现存的 Linux kernel 来提供一个集成的虚拟化方法。被 KVM 控制的虚拟机器可以被看作一个进程,它能够在任何时候打开或者关闭,并且响应时间非常短。通过使用 KVM 解决方案,当服务器休眠时,虚拟机能够进入休眠状态。这是因为 KVM 虚拟机作为一个进程,接受 Linux 的电源管理。KVM 将 Linux 的内核包含在内,它能够有效地工作在 Linux 操作系统之上,没有兼容性问题。随着

^{*} 浙江省教育厅科研项目(No. Y201121189),宁波大学学科资助项目(No. XK1097, No. XKL11072)

KVM 逐渐成熟,它被越来越多地用在分布式数据管理中。

通过比较服务器虚拟化的解决方案,选择开源的 KVM 来构建服务器虚拟化环境,在 Linux 中很容易配置 KVM 的虚拟化解决方案^[9]。但是,随着服务器数量的增加,在服务器上安装和配置虚拟化软件的工作量随之增加。在分布式数据中心的服务器虚拟化配置过程中,如何快速启动多个服务器的操作系统并配置它们,是一个很大的难题。为了快速完成虚拟化环境的配置工作,并且减少服务器的管理工作量,通过网络来配置是一个最好的解决方案。

在服务器虚拟化应用中,动态迁移作为一个重要特征,需要共享存储器的支持。通常,网络文件系统(NFS)被配置用来共享存储区域,这些存储区域是 Linux 的服务器提供的。网络文件系统服务器成了分布式数据中心的 I/O 瓶颈,因为有大量的主机需要访问网络文件系统服务器。这些虚拟服务器的虚拟磁盘映像存储在网络文件系统中,导致了性能低下的输入输出操作。所以,需要一个高性能的文件系统来满足输入输出的需求。

本文的创新之处在于:使用开源的 KVM 虚拟化解决方案和集群文件系统 Lustre^[10]构建服务器虚拟化环境,极大地减少了服务器虚拟化环境的配置时间,并解决了共享存储设备的输入输出瓶颈,提高文件并行化读写速度。

2 网络启动定制的 Linux

分布式数据环境中,在多个服务器上安装操作系统并进行配置是一项很艰巨的任务。采用网络来启动服务器操作系统的方法,能够加速这些服务器的部署和调度,并减少服务器操作系统的配置和管理难度。从网络上启动服务器的 Linux 操作系统,可以采用 PXE (pre-boot execution environment) 协议来实现,PXE 协议分为客户端和服务端,PXE 客户端在网卡的 ROM 中,当计算机引导时,BIOS 把 PXE 客户端调入内存执行,并显示出命令菜单,经用户选择后,PXE 客户端将放置在远端服务器上的操作系统通过网络下载到本地运行。

PXE 协议成功运行,需要在计算机启动时,通过 DHCP(dynamic host configuration protocol) 服务器配置计算机的 IP 地址,由于是给 PXE 客户端分配 IP 地址,所以在配置 DHCP 服务器时,需要增加相应的 PXE 特有配置。PXE 客户端所在的 ROM 中,已经存在了 TFTP 客户端。PXE 客户端使用 TFTP 客户端,通过 TFTP 到 TFTP 服务器上下载所需的 Linux 内核和根文件系统。

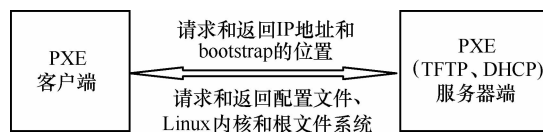


图 1 从网络启动 Linux 操作系统的过程

PXE 工作过程如图 1 所示。

PXE 客户端是需要安装 Linux 的计算机,TFTP 服务器和 DHCP 服务器运行在 PXE 服务器上,bootstrap 文件、配置文件、Linux 内核以及 Linux 根文件系统都放置在 PXE 服务器上 TFTP 服务器的根目录下。PXE 客户端在工作过程中,需要 3 个二进制文件:bootstrap、Linux 内核和 Linux 根文件系统。bootstrap 文件是可执行程序,它向用户提供简单的控制界面,并根据用户的选择,下载合适的 Linux 内核以及 Linux 根文件系统。

用于网络启动的 Linux 内核需要被定制,被定制的内核包含 3 部分内容:

- 内核必须包含服务器网络接口卡驱动。
- 内核必须支持 DHCP,内核需要从 DHCP 服务器获取一个 IP 地址。
- 当内核根据 PXE 服务器传递的参数,需要挂载网络文件系统(NFS)根目录时,内核必须支持根文件系统。可以通过配置和编译 Linux 内核源代码来产生被定制的内核。

在使用 KVM 虚拟化解决方案来构建服务器虚拟环境时,需要用到网桥功能,但是,通过传统方法从网络启动 Linux 操作系统不支持网桥功能,而网桥功能是为虚拟机分配虚拟网络接口的基本方法。网络接口控制器(NIC)作为一个网桥设备,它必须被设定为混杂模式才能够作为网桥设备来工作。通过网络启动 Linux 操作系统之后,NIC 工作在正常模式下,在 NIC 从正常模式转换到混杂模式的过程中,会有一个暂时的网络中断,这个网络中断将引起网络文件系统(NFS)根文件系统失效,将导致 Linux 操作系统的崩溃。

为了使从网络启动的 Linux 操作系统具有网桥功能,在操作系统启动过程中需要使用初始 RAM 磁盘,即 initrd。初始 RAM 磁盘是在实际根文件系统可用之前挂载到系统中的一个初始根文件系统。initrd 与内核绑定在一起,并作为内核引导过程的一部分进行加载。内核会将这个 initrd 文件作为其两阶段引导过程的一部分来加载模块,这样才能稍后使用真正的文件系统,并挂载实际的根

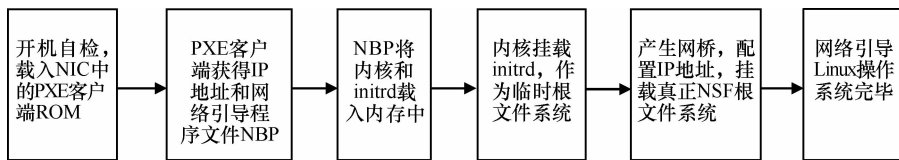


图2 initrd 网络启动步骤

文件系统。initrd 映像包含了实现这个目标所需要的目录和可执行程序的最小集合,这些可执行程序 and 系统文件用来创建网桥,并且引导 Linux 系统第二阶段的启动工作。在桌面或服务器 Linux 系统中,initrd 是一个临时的文件系统。其生存周期很短,只会用作为到真实文件系统的一个桥梁。根据不同的 Linux 版本,创建初始 RAM 磁盘的方法不相同。使用 initrd 进行网络启动的步骤如图 2 所示。

为了验证这种方法的可行性,针对 Fedora 和 Centos 操作系统进行了测试,这两个 Linux 操作系统都能够以网桥模式从网络引导并启动。KVM 软件工作稳定。

3 使用 Lustre 集群文件系统

Lustre 是工作在 Linux 操作系统之上的一个开源并行文件系统,它由 Linux 和 Clusters 演化而来,为解决海量存储问题而设计的全新文件系统,是下一代的集群文件系统,可支持 10 000 个结点,能够达到 PB 级的存储量,100 Gbit/s 的传输速度以及完美的安全性和可管理性。Lustre 提供共享访问数据,在 I/O 级和数据吞吐率方面能满足成百上千台规模的 Linux 集群服务器聚合访问的需求,数据不需要从一个文件系统复制到另一个文件系统,通过集群系统共享应用和数据的任务,执行起来要比在单独的设备上执行快得多,虚拟机可以在多个服务器之间进行动态迁移。使用 Lustre 文件系统的服务器需要安装客户端软件,并且挂载 Lustre 文件系统,KVM 的多个客户映像可以存储在 Lustre 文件系统中,在 KVM 主机之间共享虚拟服务器映像,虚拟服务器磁盘的输入输出被重定向为读写映像文件。随着映像文件数量的增多,传统的网络分布式文件系统输入输出性能成为瓶颈,而采用 Lustre 高性能并行文件系统恰好可以很好地消除此瓶颈。

由于 Lustre 集群文件系统需要修改 Linux 内核,所以

Lustre 源码在编译前必须进行修订。对于前面提到的从网络启动的服务器,它的内核必须包含 Lustre 的修订内容。在成功挂载 Lustre 文件系统后,作为 Lustre 客户端的服务器能够透明地使用共享并行文件系统来存储主机映像。图3表示出了 KVM 虚拟环境的系统结构。

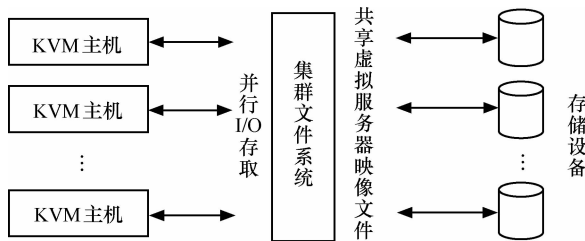


图3 KVM 虚拟环境系统结构

4 实验与分析

在使用 KVM 和 Lustre 集群文件系统的服务器虚拟环境中,针对输入输出性能进行了测试,包括普通映像文件读写、多用户大文件的读写操作。表 1 显示了包含 5 个主机的集群文件系统的文件读写测试结果。当使用单一的网络文件系统服务器作为共享存储设备时,整个虚拟环境的输入输出性能表现被单一网络文件系统服务器性能所限制;当使用 Lustre 集群文件系统作为共享存储设备时,虚拟环境的输入输出性能得到了大幅提升。使用 Lustre 集群文件系统可以解决共享存储设备的输入输出瓶颈问题。使用 Lustre 集群文件系统,20 GB 映像文件的读写时间大大减少,同时并行读写大文件的速度提高了 7~8 倍,其中,读写文件的速度为 55.1 Mbyte/s。

表 2 显示使用 Lustre 集群文件系统时,不同的结点数量情况下的各自输入输出性能对比。在集群文件系统中,结点数增长时,虽然并行读写文件的速度有所增加,性能提高 1.2~1.3 倍,但是并没有达到线性增长。

表 1 5 个主机的集群文件系统的文件读写测试结果

文件系统	20 GB 映像文件 读写时间(s)	并行读取大 文件(Mbyte/s)	并行写入大 文件(Mbyte/s)	并行读写 文件(Mbyte/s)
单一网络文件系统	365	11.5	7.5	6.8
Lustre 集群文件系统(5 台主机)	156	90.5	58	55.1

表 2 集群文件系统不同结点数的测试结果

Lustre 集群 文件系统	并行读文件 (Mbyte/s)	并行写文件 (Mbyte/s)	并行读写文件 (Mbyte/s)
5 台主机	90.5	58	55.1
10 台主机	117.6	71.3	73.2
性能提高倍数	1.3	1.2	1.3

5 结束语

在网络中启动定制的 Linux 操作系统,快速地对服务器虚拟环境进行配置,可以方便地进行服务器虚拟集成,建立服务器虚拟环境。通过网络启动 KVM 主机,可以简化 Linux 操作系统的安装和配置,减少针对多个服务器的大规模管理工作。将 KVM 服务器虚拟化解决方案和 Lustre 集群文件系统相结合,构建开源的服务器虚拟环境,可以实现服务器的动态迁移特性,虚拟环境的输入输出性能得到大幅提升,同时使得创建 KVM 用户的耗时间明显缩短。从实验结果看出,构建的 KVM 和 Lustre 虚拟环境在读写映像文件和并行读写大文件时,性能有了显著提高。当集群文件系统结点数规模增加时,性能并没有达到线性增长,在将来的工作中,可以对集群文件系统性能达到峰值时的结点数规模开展测试工作。

参考文献

1 Neiger G, Santoni A, Leung F, *et al.* Intel virtualization technology: hardware support for efficient processor virtualization. Intel Technology Journal, 2006,10(3):167~177

2 Barr K, Bungale P, Newell C, *et al.* The VMware mobile

virtualization platform: is that a hypervisor in your pocket. Operating Systems Review (ACM), 2010, 44(4):124~135

3 赵华茗,李春旺,周强. 基于 XENServer 的数字图书馆云服务平台实现研究. 电信科学,2010(S1): 39~44

4 Sun Y Q, Lin J C, Qian W, *et al.* Hyper-V micro kernel architecture security analysis. Journal of Wuhan University of Technology,2010, 32(15):130~132

5 Kivity A, Kamay Y, Laor D, *et al.* Kvm: the linux virtual machine monitor. Proceedings of the Linux Symposium, Ottawa, Ontario, Canada, 2007:225~230

6 崔泽永,赵会群. 基于 KVM 的虚拟化研究及应用. 计算机技术与发展,2011,21(6):108~111

7 Adams K, Agesen O. A comparison of software and hardware techniques for x86 virtualization. Proceedings of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems, San Jose, CA, USA, 2006

8 Dahlin M, Johnson R, Krug R B, *et al.* Toward the verification of a simple hypervisor. Proceedings of 10th International Workshop on the ACL2 Theorem Prover and Its Applications (EPTCS 70), 2011: 28~45

9 Zhang H, Wu K, Li J, *et al.* Parallel file system-supported server virtual environment in data center. Proceedings of the First International Conference on Networking and Distributed Computing, Hangzhou, China, October 2010

10 Zhao T, March V, Dong S, *et al.* Evaluation of a performance model of luster file system. Proceedings of 5th Annual ChinaGrid Conference(ChinaGrid 2010), Guangzhou,Guangdong, China, July 2010: 191~196

【作者简介】 郭延辉,宁波大学副教授,主要研究方向为并行计算和大规模分布式数据处理;陈征,高级工程师,主要研究方向为大规模分布式数据处理;龚松春,宁波大学讲师,主要研究方向为大规模分布式数据处理。

Research and Implementation of Virtual Platform on Cluster File System

Wu Yanhui, Chen Zheng, Gong Songchun

(Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China)

Abstract Virtual server environment plays an important role in managing distributed files system. It can finish large scale computing tasks, cut down hardware expenses, lessen consumption of space and power by integrating several virtual servers. Server virtualization technology can make a virtual server to migrate from one machine to another at run time. This important feature of server virtualization is dynamic migration that is realized by shared storage. With dynamic migration, many servers' planned hardware maintenance which causes service interruptions can be avoided. When the numbers of virtual servers increase, shared storage performance becomes system bottleneck. In this paper, a virtual server environment based on Lustre cluster file system is build to realize shared storage of server virtualization. Experiment results show that the virtual server's I/O performance gets remarkable promotion and less time consuming.

Key words cluster file system, server virtualization, kernel-based virtual machine

(收稿日期:2012-07-15)