第34卷 第23期 计 算 机 工 程 2008年12月
Vol.34 No.23 Computer Engineering December 2008
·博士论文· 文章编号：1000—3428(2008)23—0001—03 文献标识码：A 中图分类号：TP311

轻量级虚拟机软件技术——LVMM

沈玉良[1,2]，许 鲁[1]
(1. 中国科学院计算技术研究所，北京 100080；2. 中国科学院研究生院，北京 100039)

摘 要：为提高PC系统的可管理性和安全性，提出一种轻量级虚拟机软件技术——LVMM。定义活跃用户域，可直接访问除磁盘和网络之外的物理设备，以及虚拟磁盘和虚拟网络设备。保证在保持PC使用模式基本不变的前提下，可在同一平台上同时运行多个用户虚拟系统，且支持独立于用户操作系统的资源访问控制。经测试，LVMM原型系统具有较小的整体虚拟化开销。

关键词：虚拟机软件；服务部署；虚拟化

LVMM：A Light Weight Virtual Machine Monitor Technology

SHEN Yu-liang[1,2], XU Lu[1]
(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080;
2. Graduate University of Chinese Academy of Sciences, Beijing 100039)

【Abstract】This paper presents a light weight virtual machine monitor——LVMM for improving the manageability and security of PCs. Compared with other virtual machine system, LVMM defines the active user domain which directly accesses the hardware devices except the physical disk and network devices. It accesses the virtual disk and virtual network card devices provided by the service domain. The user mode and experience of common PC users are kept unchanged while running several user domains on the LVMM platform simultaneously. The monitoring and access controlling of system resource in LVMM is independent of the user OS. Test result shows the prototype of the light weight virtual machine monitor has a little integral virtualization overhead.

【Keywords】virtual machine monitor; service deployment; virtualization

1 概述

随着多核CPU、大容量内存、高带宽网络和高性能存储技术的不断发展，计算机的硬件资源相对于单个PC机用户的使用需求呈现出相对过剩的态势。因此，系统结构设计的主要矛盾已经从资源供给转化为如何更好地管理和使用资源，如提高计算机系统的可管理性、安全性等。然而，当前用户操作系统直接运行在硬件平台之上的系统结构使系统管理和安全程序必须依赖于用户操作系统。近年来，随着带有虚拟化支持的硬件平台和高效率虚拟化软件的进一步发展，在硬件平台和操作系统之间插入虚拟机软件，使得通过虚拟机软件进行系统平台管理和控制变为可能。

本文提出在保持PC用户使用模式基本不变的情况下，可对用户操作系统进行全生命周期管理和控制的轻量级虚拟机软件——LVMM。LVMM为实现独立于用户操作系统的安全访问控制、服务部署[1]等应用提供了技术基础。目前有多种系统级虚拟机软件，如VMware ESX[2], Xen[3]等，其主要应用目标是企业数据中心的服务器整合和资源管理。与这些通用虚拟机技术相比，LVMM有如下特点：

(1)应用目标：通用虚拟机软件技术的主要应用目标是实现服务器整合和提高资源利用率。而LVMM的主要应用目标是在PC用户使用模式基本不变的前提下，实现独立于用户操作系统的资源管理控制和提高PC用户平台的可管理性。

(2)用户域：通用虚拟机软件的用户域系统使用由虚拟机服务域提供的虚拟设备。而LVMM将第1个用户域称为活跃用户域。活跃用户域使用除磁盘和网络之外的物理设备、虚拟磁盘和虚拟网络设备。在活跃用户域中，用户还可以启动普通用户域。普通用户域完全使用由服务域提供的虚拟设备。

(3)资源虚拟化的处理方式：由于活跃用户域使用和控制平台物理设备，又要访问虚拟磁盘和虚拟网络设备，因此LVMM必须在为活跃用户域提供统一I/O资源空间的同时，采用不同的I/O处理方式。此外，LVMM还对活跃用户域的内存映射、I/O访问等方面进行了优化处理。

2 背景介绍

虚拟机系统一般具有一层介于物理平台和用户操作系统间的虚拟机软件(Virtual Machine Monitor, VMM)。虚拟机软件负责管理对物理平台的访问，以使物理平台能够被多个用户虚拟系统共享使用。

由于早期设计的x86平台没有考虑虚拟化的问题，因此在传统的硬件平台上实现虚拟机软件需要一定的复杂技巧来绕过x86平台的一些特性。近期，Intel和AMD先后推出了各自硬件平台虚拟化的扩展技术：Intel的VMX[4]和AMD的SVM[5]。硬件虚拟化技术的出现降低了在x86平台实现虚拟机软件的难度，也进一步推动了虚拟机软件应用的发展。

目前，虚拟机软件在服务器整合、软件调试和测试等方面的应用取得了重要发展。与此同时，各大公司和研究机构都在探索虚拟机技术在其他领域的应用。例如：Intel公司开始研发的EIT(Embedded IT)架构也是希望在用户操作系统运

基金项目：国家“973”计划基金资助项目(2004CB318205)
作者简介：沈玉良(1977－)，男，博士研究生，主研方向：服务部署，虚拟机技术；许 鲁，研究员、博士、博士生导师
收稿日期：2008-04-15 **E-mail：**shenyuliang@nrchpc.ac.cn

行的同时，运行一些虚拟系统(Embedded OS)来提高企业 IT 设施的可管理性。联想等 PC 厂商也纷纷研究如何利用虚拟机技术提高 PC 系统的安全性能。

3 轻量级虚拟机技术

与通用虚拟机技术相比，LVMM 在域的定义、硬件平台共享机制、内存管理策略、I/O 资源虚拟化等方面有不同的设计思路和实现方式。

3.1 轻量级虚拟机系统结构

LVMM 定义了 3 种类型的域：服务域，活跃用户域和普通用户域。服务域是提供系统服务的域，可以直接访问硬件平台上的磁盘和网络设备，并负责用户虚拟系统的管理和 I/O 设备虚拟化；活跃用户域是系统启动的第 1 个用户域，并运行标准的用户操作系统。它直接访问硬件平台上除磁盘和网卡的物理设备，如显卡、USB 设备等，并使用由服务域提供的虚拟磁盘和网络设备。因此，活跃用户域是用户直接使用的计算系统；其后启动的用户虚拟系统，称为普通用户域。它们访问和使用的 I/O 设备都是由服务域提供的虚拟设备。

LVMM 的系统结构如图 1 所示。

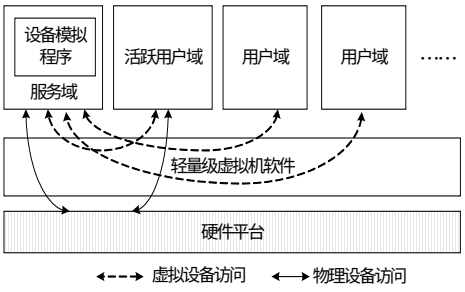

图 1 LVMM 的系统结构

3.2 内存资源虚拟化

为优化 LVMM 的内存虚拟化处理机制，本文为活跃用户域增加了直接内存映射机制，即把活跃用户域的大部分物理地址直接映射到硬件平台机器地址。内存地址的直接映射既可简化虚拟机软件内的地址映射处理，又可使活跃用户域物理设备的 DMA 地址等于硬件平台机器地址。

LVMM 在初始化时会检测硬件平台的物理内存大小，并将 LVMM 自身的内存堆和服务域映射到硬件平台内存的高端地址。同时，将活跃用户域 16 MB 以上的物理内存映射到直接对应的硬件平台机器内存地址空间，并将 1 MB~16 MB (虚拟机程序除外)的内存直接映射到硬件平台内存地址空间。活跃用户域的 VGA RAM(0xA0000-0xC0000)也会被直接映射到硬件平台内存地址空间。在活跃域中小于 1 MB 的其他物理内存空间会被映射到由服务域提供的虚拟系统 BIOS。在活跃域地址空间中对应 LVMM 程序占用空间的部分会被映射到高端内存区域。内存映射关系如图 2 所示。

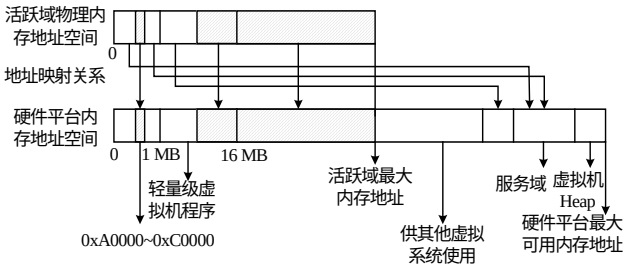

图 2 LVMM 的内存映射

3.3 I/O 资源虚拟化

LVMM 将大部分物理 I/O 资源赋给活跃用户域直接访问，同时为其提供虚拟磁盘和网络设备。因此，LVMM 设计和实现了与通用虚拟机软件不同的 I/O 虚拟化处理机制。

3.3.1 I/O 指令虚拟化

LVMM 针对不同设备的 I/O 指令采用不同的处理方式，具体处理机制如图 3 所示。

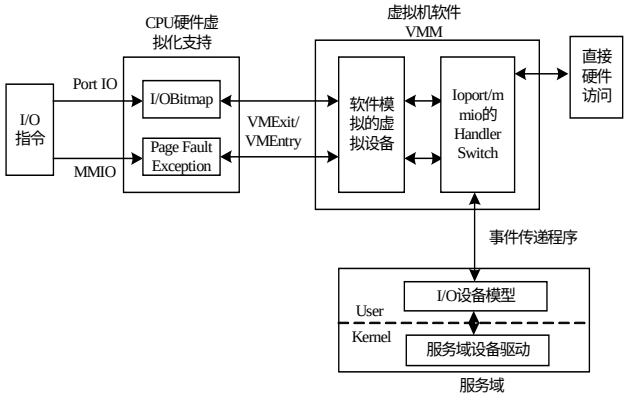

图 3 I/O 指令虚拟化机制

活跃用户域执行的 I/O 指令可分为 Port I/O 和 Memory Map I/O。对于 Port I/O，系统硬件会根据 CPU 的 IOBitmap 的设置情况决定是否触发 VM Exit。如果 IOBitmap 中对应端口的位值为 0，则直接执行该 I/O 指令。否则，系统执行 VM Exit，并进行 I/O 指令模拟。对于 MMIO，LVMM 会把 MMIO 地址区域的页表属性设置为缺页。当系统执行相关指令时，系统会产生 Page-Fault 异常，并导致触发 VM Exit。此时，LVMM 会检测导致 VM Exit 的原因，并进行 I/O 请求处理。如果该 I/O 指令访问的地址范围属于 LVMM 内部的虚拟设备，则 LVMM 会处理该 I/O 请求。否则，LVMM 会把该 I/O 请求交给 switch 程序判断是否直接访问物理设备。如果是，则由 LVMM 直接进行硬件平台访问。否则，LVMM 会把该 I/O 请求发送给服务域进行处理。服务域的虚拟设备程序可以根据 I/O 请求内容，执行相关处理，并把结果返回给 LVMM。LVMM 根据 I/O 处理结果，设置相关寄存器的内容，并执行 VM Entry 操作继续用户操作系统的执行。

综上所述，LVMM 采用从物理硬件、虚拟机软件到服务域虚拟设备程序的多级 I/O 处理机制，并依此实现了活跃用户域的物理设备和虚拟设备混合的 I/O 空间。

3.3.2 PCI 设备的虚拟化

LVMM 为活跃用户域提供了虚拟的 PCI 配置空间。在物理 PCI 总线 0 上添加了软件虚拟的 Intel 21152 PCI 桥设备。在虚拟 PCI 桥的另一端，为活跃用户域添加了虚拟的 PCI-IDE 设备和虚拟网卡。

对于虚拟 PCI 配置空间的访问，LVMM 会检查该请求是否针对物理 PCI 设备。如果访问的是物理设备，则检查内部缓存的 PCI 配置空间信息。如果缓存命中，则进行相关信息的读写。如果未命中，则根据策略访问物理 PCI 配置空间或返回默认值。如果访问的是虚拟设备，LVMM 会把该请求转发给服务域内的设备虚拟程序进行处理。LVMM 实现 PCI 配置空间缓存机制的好处在于可有效控制活跃用户域对物理 PCI 配置空间的访问，并通过 PCI 设备的基址寄存器(BAR)控制其 I/O 地址空间的映射。

3.3.3　设备中断的虚拟化

为了能让 LVMM 控制硬件平台的物理中断处理，本文为活跃用户域提供了虚拟中断控制器。这样，LVMM 和活跃用户域有不同的中断向量空间。对于活跃用户域使用的物理设备和虚拟设备，LVMM 采用不同的中断处理方式。

对于虚拟设备，服务域的设备虚拟化程序发送中断请求事件给 LVMM。然后，由 LVMM 利用活跃用户域的虚拟中断控制器产生设备中断。

当物理设备产生中断时，LVMM 进行中断虚拟化的大致处理步骤如下：

(1)LVMM 将所有和硬件平台设备相关的中断向量的处理函数都设置成 *do_IRQ* 函数；

(2)当中断发生时，*do_IRQ* 函数先认可该中断向量；

(3)*do_IRQ* 函数检查该中断向量是否分配给用户域系统；

(4)如果该中断向量没有分配给用户域或服务域系统，则 LVMM 直接调用内部的中断处理逻辑进行处理；

(5)否则，*do_IRQ* 函数会检查与相应中断线关联的设备是否全部属于服务域；

(6)如果与该中断线关联的设备全部属于服务域，则 LVMM 向服务域发送中断信号事件来调用服务域的中断处理函数进行处理；

(7)如果与该中断线关联的设备全部属于活跃用户域，则 LVMM 通过 CPU 硬件虚拟化机制向活跃用户域插入相应的中断请求，并设置虚拟中断控制器；

(8)如果与中断线关联的设备既包括服务域设备，又包括活跃用户域设备，即服务域设备和活跃域设备共享该中断线，LVMM 将该中断请求先交付给优先级高的域进行处理；

(9)如果高优先级域成功地处理了中断请求(如发送 EOI)，则清 0 失败记数值(*Fcount*)，并跳至(12)；

(10)如果高优先级域在规定时间内，仍没有成功处理该中断，则 *Fcount* 加 1，并把该中断请求发送给低优先级域进行处理；

(11)如果 *Fcount* 大于 3，则调整服务域设备和活跃用户域设备的优先级；

(12)中断处理结束。

由上述 LVMM 的中断处理过程可以看出，在服务域设备和活跃域设备共享中断线时，系统会产生较长的中断延迟。为避免由此引入的性能损失，LVMM 的设备配置程序内采取了尽量避免 2 个高速 I/O 设备跨域共享中断线的策略。

3.3.4　物理设备的 DMA 处理机制

用户操作系统的物理地址空间，须经过虚拟机软件的地址转换机制才能翻译成真实的机器内存地址空间。在支持 I/O 虚拟化的硬件平台中，系统 I/O 控制器可以根据虚拟机软件提供页表来自动进行用户操作系统的物理地址到真实机器地址的翻译。这样，物理设备就能正确地进行 DMA 操作。对于不支持 I/O 虚拟化的系统平台，须进行复杂的虚拟化处理。在 LVMM 中只有活跃用户域、服务域操作和使用物理设备。本文对服务域操作系统内核中关于 DMA 操作的部分进行了修改，使其能在完成内存地址转换后再进行 DMA 操作。在活跃用户域中，绝大部分物理地址和硬件机器地址一致。因此，LVMM 无须对活跃用户域进行复杂的 I/O 地址翻译操作。

4　性能分析和潜在应用

4.1　轻量级虚拟机性能分析

目前笔者已完成 LVMM 原型系统的实现，对 LVMM, Xen 和本地操作系统的性能进行了比较测试，见图 4。

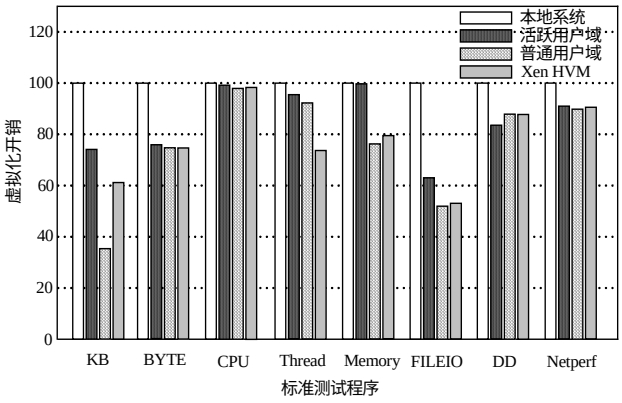

图 4　轻量级虚拟机系统性能比较

本次测试采用 Intel 965P 平台，Core2 1.86 GHz CPU, 1 GB DDR2 667 内存，希捷 120 GB 硬盘和 Marvell 88E8053 网卡，操作系统是 Fedora Core 5。活跃用户域配置为双虚拟 CPU、512 MB 内存、Realtek 8139 网卡和 20 GB 硬盘。普通用户域被配置成 256 MB 内存，其他配置与活跃用户域一致。本次性能测试采用的标准测试程序有内核编译(KB)、BYTE、Sysbench(包括 CPU、Thread、Memory、FILEIO)、磁盘 DD 以及 Netperf，同时把与活跃用户域配置相同的 Xen 3.1 HVM 用户域作为参考对比。为了比较轻量级虚拟机系统与本地系统的性能差距，将本地操作系统的性能测试结果定义为基准值 100。

通过上述测试结果，可以看出活跃用户域和普通用户域在 CPU 和内存方面的虚拟化开销较小。而在磁盘、文件读写、网络等相关方面的虚拟化开销较大。LVMM 和 Xen 相比具有较小的虚拟化开销。这主要是由于 LVMM 在活跃用户域的内存虚拟化处理中采用了直接内存映射的优化方案。

4.2　轻量级虚拟机的潜在应用

本文以服务部署系统和安全保护系统为例，探讨 LVMM 在桌面系统的管理和安全保护等方面的潜在应用。

(1)服务部署系统

服务部署系统是一种基于集中存储的计算机部署管理系统。其部署机制能将物理计算资源与存储资源快速、灵活地结合为可运行的计算系统供用户使用。然而，以软件方式实现的部署机制会依赖于用户操作系统。面对不同种类的用户操作系统会有相应的部署程序移植。如果基于 LVMM 实现系统部署，则只需要在 LVMM 中实现 1 次，就能够适用于标准操作系统的部署。这是由于 LVMM 为活跃用户域提供的是标准磁盘访问接口和标准的 PCI 设备。此外，LVMM 还可方便地实现磁盘和网络设备的运行时监控。

(2)安全保护系统

安全保护系统主要是指数据保护、反病毒等系统安全应用程序。对于安全保护程序来说，需要备份恢复系统的近期修改或清除恶意程序。如果由于用户或程序误操作等原因，有可能导致系统重要文件损坏。这时，操作系统无法启动，系统安全应用程序更无法运行。如果基于 LVMM 实现相关的系统安全程序，就可以在用户操作系统启动之前运行系统安全保护程序。这表示可在用户操作系统启动之前，恢复误删除的数据或进行数据安全检测。

（下转第 19 页）