◎研发、设计、测试◎

文件支持的 Xen 存储虚拟化研究

汤 泉,李小勇
TANG Quan,LI Xiao-yong

上海交通大学 信息安全学院,上海 200240
Department of Information Security,Shanghai Jiao Tong University,Shanghai 200240,China
E-mail:qdtangquan@hotmail.com

TANG Quan,LI Xiao-yong.Research of file-backed Xen storage virtualization.Computer Engineering and Applications, 2009,45(16):77-79.

Abstract: File-backed virtual disk storage is one of the most important parts of virtualization. In order to enhance the efficiency of virtual disk I/O,this paper researched the file-backed disk virtualization on Xen,analysed using asynchronous I/O method to improve I/O rate,and compared the pros and cons of using different format of virtual disk files as storage device of virtual machines.

Key words: storage virtualization;virtual disk image;Xen

摘 要: 基于文件的虚拟磁盘存储是虚拟机技术实现的重要一环,为了提高虚拟机磁盘读写效率,着重研究了 Xen 基于文件的磁盘虚拟化,分析利用异步 I/O 技术提升虚拟机对虚拟磁盘的读写速度,同时阐述了不同格式的虚拟磁盘文件做为虚拟存储设备的优劣。

关键词: 存储虚拟化;虚拟磁盘镜像;Xen
DOI: 10.3778/j.issn.1002-8331.2009.16.021 **文章编号:** 1002-8331(2009)16-0077-03 **文献标识码:** A **中图分类号:** TP302

虚拟化技术源于大型机,1972 年,IBM 在其 S360 大型机上首次实现虚拟机模型[1],它允许多个系统运行在一台机器上,提高了服务器的利用效率,距今已经有近 40 年的历史了。互联网发展起来以后,新兴的虚拟机应用不断出现,使虚拟化技术成为当前比较热的一个研究领域,出现了许多比较成熟的各种虚拟机或模拟器技术和产品,如 Xen、Vmware 和 Qemu,而存储虚拟化则是实现它们的一个很重要方面。

这里的存储虚拟化特指虚拟机中的虚拟存储设备,它与传统意义上存储设备所要实现的目标是一致的,需要简化存储管理系统,提高存储利用率和性能,增强存储数据安全性。本文着重研究了单机上基于文件的 Xen 磁盘虚拟化原理, 分析利用异步 I/O 技术提升虚拟机对虚拟磁盘的读写速度,同时从理论和实验两方面说明了使用 Raw 格式和 Qcow 格式虚拟磁盘文件做为虚拟存储设备在性能和功能上的差异。

1 相关背景

1.1 Xen

Xen[2]是剑桥大学开发的一款开源 X86 虚拟机管理器,其上可以同时运行 100 多个不同的操作系统,它采用了准虚拟化(Para-virtualization)的方法,通过少量修改虚拟机中的内核,让其主动与虚拟机监视器(VMM,Virtual Machine Monitor)协同工作来达到很高的性能。实验表明,Xen 中虚拟机的性能和真实的机器性能相差在 3%左右[2]。从 Xen3.0 版本开始,Xen 加入了对 Intel VT 和 AMD pacifica 技术的支持,大大提升了 VMM 对虚拟机的掌控能力,实现了全虚拟化(Full-virtualization),可以运行不加修改的操作系统,同时也考虑到了如何有效减小虚拟机的开销[3]。

在 Xen 中,Domain 指一个运行中的虚拟机,在引导时最初创建的 Domain 称为 Domain 0,它被允许使用控制接口来操控应用级的管理软件,可以创建和终止其他 Domain,控制它们相关的调度参数、物理存储分配以及对给定的存储设备或网络设备的访问。

1.2 存储虚拟化

新存储实体对原存储实体的存储资源和存储管理进行变化和转换的过程称为存储虚拟化,存储资源包括存储的读写方式、连接方式、存储的规格或结构等,而存储管理包括统一管理、分散管理、性能动态调整管理等。从虚拟机的角度来看操作的是虚拟设备,而不必关心真正的物理设备是什么。

为虚拟机提供的虚拟存储设备可以有以下几种方式[4]:物理硬盘或其分区,包括 iSCSI 磁盘或 GNBD 卷;网络存储

基金项目: 国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z169)。
作者简介: 汤泉(1984-),男,硕士研究生,主要研究领域为网络安全,虚拟机;李小勇(1972-),男,博士,副教授,硕士生导师,主要研究领域为高性能网络、虚拟计算、嵌入式系统、网络与信息安全技术等。
收稿日期: 2008-04-14 **修回日期:** 2008-07-03

协议,包括 NFS 等网络或并行文件系统;基于逻辑卷(Logical Volumn Manager,LVM);基于文件的虚拟存储设备,也称为虚拟磁盘镜像(Virtual Disk Image)。本文重点讨论的就是最后一种,基于文件的虚拟存储方式。

1.3 虚拟磁盘文件

虚拟磁盘文件顾名思义就是以一定格式的文件来做为虚拟机的磁盘存储设备,这种方式的优点是个人用户配置灵活、使用方便。每种虚拟机技术几乎都有与之相对应的虚拟磁盘文件格式:最原始的是 Raw 格式,它是一种"直读直写"的格式,不具备特殊的特性,容易被其他程序所读,Linux 中直接可以以回环(loop)设备来将它挂载到一个目录下;Qcow 格式是 Qemu 中实现的一种镜像文件格式;VMware 使用的是 Vmdk 格式;微软在其虚拟个人电脑和虚拟服务器上使用的 Vhd 格式等等。本文重点讨论的是在 Xen 虚拟机上使用 Raw 和 Qcow 格式的虚拟磁盘文件。

2 基于文件的 Xen 磁盘虚拟化模型设计

2.1 磁盘虚拟化原理

在准虚拟化的 Xen 中,虚拟存储设备由 Domain 0 负责创建和赋给其他 Domain,只有 Domain 0 可以使用 Linux 驱动程序对真实磁盘进行直接的访问,其他的由 Domain 0 创建的 Domain U 只能借助于 VMM 和 Domain 0 来访问磁盘设备。具体实现方法是:在 Domain 0 经修改后的内核里添加了一个后端虚拟驱动(Back-end Driver),为其他虚拟机提供对网络和块设备的访问,与此对应,在其他的准虚拟化 Domain U 的内核中添加了一个前端虚拟驱动(Front-end Driver)。

如图 1 所示,前端驱动位于虚拟机 Domain U 中,接受来自 Domain U 内核的 I/O 请求,传递给后端驱动,后端驱动位于 Domain 0 中,负责接收来自前端的 I/O 请求,因为两者位于不同的内核中,所以它们之间的通信要依赖 VMM 中的 I/O 共享环和事件通道。Domain U 将 I/O 请求放到环上并移动请求生产者指针,Domain 0 将这些请求移出环处理,并移动相对应的请求消费者指针;经处理后生成的响应被放回到相同的环上并移动应答生产者指针,同时等待 Domain U 的前端驱动读取和移动应答消费者指针。另外 Xen 使用事件通道作为有 I/O 描述符进入队列的异步通知,不管是请求还是应答,都可以在环里

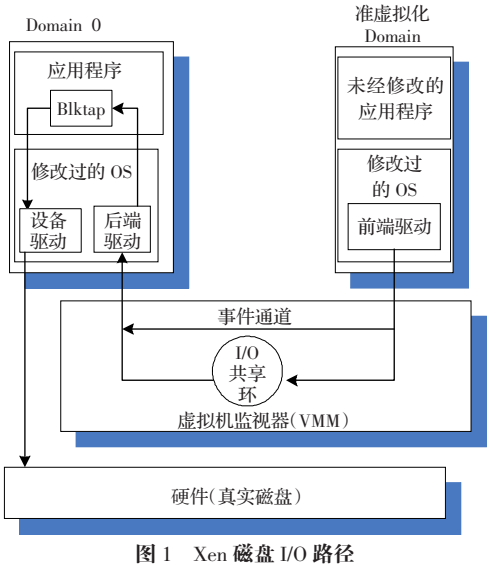

图 1 Xen 磁盘 I/O 路径

同时放入多个它们的描述符项,直到达到一定的阀值后才发送事件通知[5]。当 Domain 0 的后端驱动得到传来的 I/O 请求后,通过字符设备将请求送至用户态程序 Blktap,经其处理和转换后才调用 Domain 0 内 Linux 操作系统中真正的磁盘设备驱动程序最终完成对物理磁盘的访问请求。

2.2 Blktap 功能

Blktap 是一个运行在 Domain 0 用户空间的程序,给用户层提供对虚拟磁盘文件读写操作的接口。如图 2 所示,比如某个 Domain 所对应的虚拟磁盘设备是真实磁盘上的文件 A,则 Blktap 的主要作用就是将此 Domain 所要访问的虚拟磁盘 64 位扇区号转化为对物理磁盘上文件 A 中相应偏移量的操作。

Blktap 有很好的可扩展性,可以很容易实现支持不同格式的虚拟磁盘文件做为虚拟机的虚拟存储设备,同时易于实现如写时拷贝、磁盘加密、磁盘压缩等不同的应用,一些已有的块资源管理和磁盘调度策略也能通过它方便地附着在块设备上,另外因为 Blktap 处于用户空间,所以一些用户层的库和工具可以方便地被其调用以提高性能。

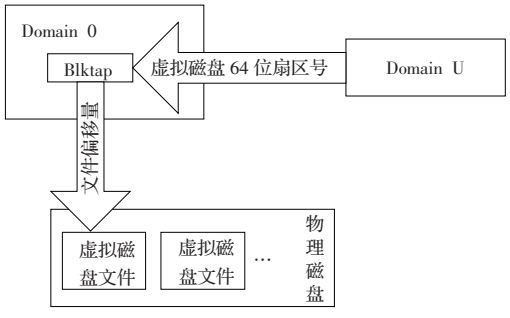

图 2 Blktap 作用

2.3 加入异步 I/O 机制

Blktap 原有的读写模式是同步的,当它把从虚拟机传来的某扇区号的访问转化为对实际的文件某偏移处的操作,执行一个系统调用时会导致 Blktap 阻塞,直到系统调用完成即数据传输完成或发生错误为止,显然阻塞期间,CPU 一直处于空闲状态,没有得到充分的利用。而如上所述,用户空间的 Blktap 能方便地使用已有的工具库,于是很自然地想到对其使用 Domain 0 的 Linux 操作系统自带的异步读写库。利用处理速度与 I/O 速度之间的差异,当一个或多个 I/O 请求挂起时,CPU 可以执行其他任务,或者在发起其他 I/O 的同时对已经完成的 I/O 进行操作,增加了 CPU 的利用率,相应也提高了对虚拟机的虚拟磁盘设备的读写效率。异步处理流程如图 3 所示。

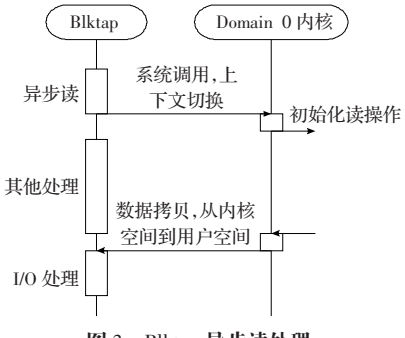

图 3 Blktap 异步读处理

2.4 虚拟磁盘文件格式

基于文件的虚拟存储设备可以由多种不同格式的文件来

实现,加上了 Blktap 的支持后使这种实现和扩展更为容易,这里着重列出 Xen 常用的 Raw 和 Qcow 两种虚拟磁盘文件格式。

Raw 格式是一种"直读直写"的格式,Raw 格式虚拟出的磁盘中,磁盘块号越大,此块在文件中所处的偏移量越大,即文件中数据存放的排列顺序与虚拟出磁盘上的数据块顺序相一致,因此从磁盘号到偏移量的地址转换比较方便快捷,但是缺少了元数据区域的支持,容易导致文件系统和数据拷贝出错。

Qcow 格式是 Qemu(动态二进制翻译处理器模拟器)中实现的一种镜像文件格式。在 Qcow 格式文件中,数据存储的基本单元是 cluster[6]。每个 cluster 都由几个块组成,每个块占 512 个字节,要访问某个特定的 cluster 需要经过两次查询操作,这个过程有点类似内存二级页表查询的机制。每个由虚拟机传来的 64 位虚拟磁盘扇区号按 cluster_bits 和 l2_bits 两个字段的大小被分成 3 个部分,举个例子:如果 cluster_bits 是 12、l2_bits 是 9,则一个给定地址的最低 12 位标识要访问的块在某个特定 cluster 内的偏移量,次低的 9 位标识一个 512 项数组内的偏移量,它指向那个特定的 cluster,剩余的 43 位标识在一级表数组内的偏移量,它指向刚才的那个二级表,当一级表或者二级表某表项记载的偏移量为 0 时表示相应的地址还没有在映像文件中被分配,所以 Qcow 格式的文件可以动态增大。

Qcow 格式文件起始处有专门的区域存放元数据,支持一些高级的特性,如加密、压缩和快照,以节省磁盘空间并保证数据安全。另外与 Raw 格式相比最重要的优势在于 Raw 文件需要在创建时就分配好足够大的空间,而 Qcow 文件可以从一开始的小文件然后随着内容的增多而动态地逐渐增大,这种实现大大提升了对真实物理磁盘的利用率。

3　性能评估

3.1　实验环境

实验机器配置如下:CPU:Intel Pentium 4 3.0 GHz,内存:1 G,磁盘:60 G。装上 Xen 3.0.4 的 Fedora 6(内核 2.6.16.33 Xen0),在其上运行一个 Fedora6(内核 2.6.16.33 XenU)的虚拟机,分配给虚拟机的内存大小为 256 M。总共测试 3 次,前两次此虚拟机都是使用 Raw 格式镜像做为虚拟磁盘,只是第一次测试时使用的虚拟磁盘读写操作是同步的,第二次则加入了异步机制,最后一次测试使用 Qcow 格式镜像做为虚拟磁盘,读写方式为异步。所有的测试结果都是在虚拟机中运行开源的磁盘 I/O 测试工具 Bonnie 得到,测试文件大小一般为 2 倍使用的内存,所以设定测试文件为 512 M。

3.2　同步异步性能比较

在基于 Raw 格式虚拟磁盘的虚拟机上,分别对用 Blktap 同步和异步两种方式访问虚拟磁盘进行测试,实验结果如图 4 所示,Bonnie 测得关于对虚拟磁盘的各项 I/O 指标,异步比同步效率高出 6%到 16%不等,这是因为对 Blktap 加入异步读写机制后,CPU 利用率提升了,实验结果与理论分析相一致。

3.3　使用不同文件格式性能比较

新建两个操作系统完全一致的虚拟磁盘文件,一个是 Raw 格式,另一个是 Qcow 格式,在同一台主机上分别运行这两个格式不一致、内容完全一致的虚拟机,然后在两台虚拟机里分别运行 Bonnie 测试程序,实验结果如图 5 所示。从图中可以看

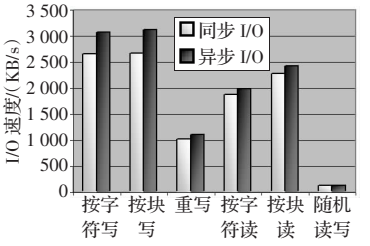

图 4　虚拟磁盘同步与异步 I/O 结果比较

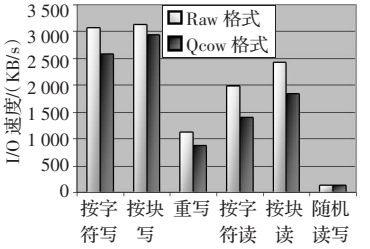

图 5　读写 Raw 格式与 Qcow 格式文件性能比较

出使用 Qcow 格式文件在读写各项指标上都稍逊于使用 Raw 格式,这是由于 Qcow 的偏移量寻址需要通过二级查表,另外 Qcow 格式的虚拟磁盘是动态增长的,在写磁盘时经常要去申请新的空间,这些是在性能上稍慢的原因。但是以时间代价换来了可加密、压缩、快照等功能,以及磁盘空间的节省和读写安全性的提升,因此 Qcow 仍是 Xen 用于虚拟磁盘文件格式的首选。

4　结束语

本文通过研究 Xen 的 Blktap 模块,分析了虚拟机实现基于文件的虚拟磁盘设备的大致原理,在其上加入异步磁盘读写机制后,显著提升了虚拟机对虚拟磁盘的读写效率。另外,使用 Qcow 格式虚拟磁盘文件虽然在读写效率上比 Raw 格式有所下降,但它提供了可动态增长、可加密、可压缩和快照等优越性能,将是 Xen 所使用的主要虚拟磁盘文件格式,而且针对它的独特的寻址机制,可考虑加入快表和缓存技术,应该还有更大的性能提升的空间。

参考文献:

[1] Creasy R J.The origin of the VM/370 time-sharing system[J].IBM Journal of Research and Development,1981.

[2] Barham P,Dragovic B,Fraser K,et al.Xen and the art of virtualization[C]//19th ACM Symposium on Operating Systems Principles,Oct 2003.

[3] Pratt I,Fraser K,Hand S,et al.Xen 3.0 and the art of virtualization[C]//Proc of the 2005 Ottawa Linux Symposium,Ottawa,Canada,July 2005.

[4] Xen 3.0 user manual.http://www.cl.cam.ac.uk/research/srg/netos/xen/readmes/user.pdf.

[5] Fraser K.Safe hardware access with the Xen virtual machine monitor[C]//OASIS ASPLOS 2004 Workshop,Boston,Mass,October 9,2004.

[6] McLoughlin M.The QCOW image format.http://www.gnome.org/~markmc/qcow-image-format.html.