

一种基于 iSCSI 协议的动态扩容机制

田 龙, 张延园, 张 晓

(西北工业大学 计算机学院, 陕西 西安 710129)

摘 要: 为了解决海量存储系统研究过程中 PB 级硬件需求过高的问题, 实现了一种基于 iSCSI 协议的容量仿真技术, 通过分析文件系统的格式, 进一步提出了更具通用性的地址映射方案和动态扩容机制. 目前 iSCSI 提供了远程存储机制, 在此基础上使用一定大小的文件模拟磁盘, 并对磁盘容量进行虚拟扩大和地址映射管理, 实现了小文件到大磁盘的仿真, 同时对动态地址映射方案下的动态扩容机制予以介绍. 最后通过实验对虚拟磁盘和普通磁盘的性能进行了测试和对比, 证明了该技术的可行性和正确性.

关键词: 容量仿真; iSCSI 协议; 地址映射; 动态扩容

中图分类号: TP391.9

文献标识码: A

文章编号: 1000-7180(2011)06-0058-04

Dynamic Capacity Expansion Mechanism Based on iSCSI Protocol

TIAN Long, ZHANG Yan-yuan, ZHANG Xiao

(College of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China)

Abstract: In order to solve the problem of PB grade hardware demand in the mass storage system research process, a capacity simulation technology based on iSCSI protocol was proposed. Through the analysis of the file system format, a kind of more general mapping scheme and dynamic capacity mechanism were put forward further. At present, iSCSI provides a remote storage mechanism, on the basis of which we simulate disks using certain size files. The capacity are virtually enlarged and managed by address mapping, implementing the simulation from small files to large disks. At the same time, dynamic capacity expansion mechanism was introduced. At last, through experiments we test and analyze the performances between virtual disk and common disk, proving the feasibility and correctness of the technology.

Key words: capacity simulation; iSCSI protocol; address mapping; dynamic capacity expansion

1 引言

海量存储时代的到来增加了对存储资源的需求, 同时也使对 PB 级存储环境下的各方面研究和评测显得尤为必要. 该技术以国家“八六三”项目为研究背景, 旨在对 PB 级环境下集群文件系统的功能方面予以评测. 在对海量存储系统研究的过程中首先需要解决的就是海量存储环境的搭建, PB 级存储环境大大增加了其研究成本, 因此需要一种技术对所需环境进行仿真模拟, 以减少硬件环境带来的成本消耗.

iSCSI 协议提供了远程存储机制^[1], 可以把本地的磁盘提供给远程的客户端进行访问, 实现了虚拟磁盘技术. 通过分析 iSCSI 的机制, 我们实现了对磁盘的容量的虚拟, 从小文件到大磁盘的仿真, 并通过地址映射实现了对虚拟磁盘的读写管理, 实现客户对虚拟磁盘的正确访问. 在地址映射过程中, 我们提出并实现了动态的地址映射方案, 通过按写分配资源的思想使该技术能够应用于不同的文件系统, 在此基础上进一步设计并模拟实现了动态扩容机制.

最后, 我们使用磁盘性能测试工具 IOmeter 对

收稿日期: 2010-09-20; 修回日期: 2010-12-01

基金项目: 国家“八六三”计划重大项目(2009AA01A404)

虚拟磁盘和普通磁盘的性能进行了测试对比,并对结果进行了分析。

2 原理及模型

2.1 原理分析

在 iSCSI 的通信结构中由启动器 Initiator 和目标器 Target 组成^[2-3]。Initiator 端负责生成 CDB (Command Descriptor Block), 将其封装为 iSCSI PDU(Protocol Data Unit), 并通过 IP 网络将 PDU 发送给 Target。Target 端接收到 PDU 之后, 将其还原为 SCSI 命令和数据, 并将 CDB 传送给 SCSI 层并执行。

iSCSI Target 服务启动时会首先解析配置文件, 配置文件中设置了提供给客户的 lun, Target 会根据 lun 的类型通过文件操作函数 `filp_open()` 获得磁盘的操作句柄和容量大小等信息。当客户进行连接时, 服务端就会把 lun 的信息包括容量大小报告给客户, 客户端操作系统就可以根据这些信息对 lun 进行操作。

在 Target 向 Initiator 报告容量时, 我们可以主观地对磁盘大小进行更改, 进而或得比实际容量大的多的虚拟磁盘, 这样客户端操作系统就可以获得更大的寻址空间。但由于虚拟容量和实际地址映射之间无法一一对应, 因此若要完成对该虚拟磁盘的操作, 还需要一种地址映射机制对其进行管理, 以下我们将继续介绍地址映射原理。

2.2 地址映射

iSCSI 通过对 SCSI 操作的封装来实现远程的操作。在 Write 操作中, 启动器首先向目标器发送 SCSI Write 命令, 目标器根据接收到的 SCSI 命令分配缓存接收数据。在 SCSI 的命令描述块 CDB 当中, 包含了 Write 操作中所要写的数据的长度和在磁盘中的存储位置的偏移量等参数, 十字节的 SCSI 命令表 1 所示。

表 1 SCSI 十字节命令格式									
	7	6	5	4	3	2	1	0	
0	READ(28H)或 WRITE(2AH)								
1	(LUN) OPO FUA 保留 Rel								
2	逻辑块的起始地址								
3									
4									
5									
6	保留								
7	传输长度								
8									
9	控制字节								

由命令格式可知, 第三字节到第六字节用于存储读写操作的映射地址, 第八字节和第九字节用于存储所要传输数据的长度, 当目标器接收并解析 SCSI Write 命令后, 通过移位和或运算得到将要传输的数据的存储位置和长度。

计算公式如下:

$$\begin{aligned} \text{off} &= (\text{u32}) \text{cmd}[2] \ll 24 | (\text{u32}) \text{cmd}[3] \ll 16 | \\ & (\text{u32}) \text{cmd}[4] \ll 8 | (\text{u32}) \text{cmd}[5] \\ \text{len} &= (\text{cmd}[7] \ll 8) + \text{cmd}[8] \end{aligned}$$

首先通过命令的操作码来判断对数据所进行的操作, 然后通过计算获得读写操作的偏移量和数据长度, 这样就可以在命令的解释过程中对磁盘上操作的地址部分进行主观的映射, 使得读写操作可以正确执行, 进而实现对虚拟磁盘的读写管理。

2.3 iSCSI 模型结构

我们的目标是使用较小的磁盘文件模拟大容量空间, 以降低硬件资源方面的需求, 因此我们使用一定容量的文件代替分区或卷来进行虚拟操作, 在容量虚拟时采用的虚拟容量值为 2 TB, 整体模型设计如图 1 所示。

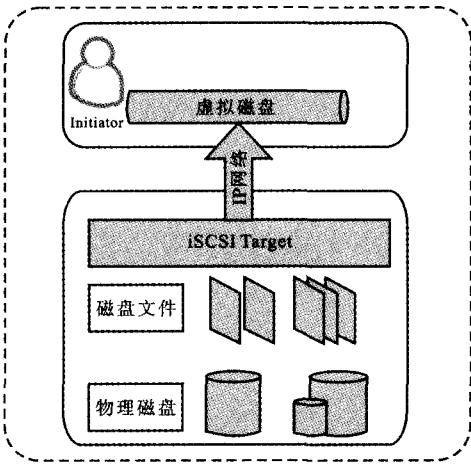


图 1 iSCSI 模型图

首先从物理磁盘上创建一定大小的分区或文件, 此处我们使用大小为 200 MB 的文件作为 lun 提供给 Initiator 客户端。在向客户端报告 lun 的同时对磁盘的容量大小进行主观设置, 使得客户端看到的是容量为 2 TB 的虚拟裸盘。客户在虚拟裸盘上创建文件系统之后, 就可以对该磁盘进行读写操作, 所以首先必须实现容量的虚拟和地址映射管理。

3 实现

3.1 大空间虚拟

iSCSI 协议中, Target 端是通过两个重要参数

向 Initiator 端报告其分区容量的. 一个是块的大小, 另外一个是个数. 其中, 块的大小采用默认值, 不能更改, 通常情况下是 512 B; 块的个数是通过磁盘索引节点的大小 (inode size) 除以块大小得到的, 可以人为的设置. 32 位的情况下, 其最大值为 0xffffffff, 转化为十进制是 4294967295, 由于块号从 0 开始, 因此最大的块号为 (0xffffffff-1). 如果采用这种设置, Target 向 Initiator 报告的分区容量为 $4294967295 \times 512 = 2199023255040$ B, 即虚拟容量最大可设置为 2 TB.

3.2 固定地址映射

由于是从小文件映射到大磁盘, 因此大磁盘中的大部分空间没有实际的物理地址映射, 地址映射的目的首先是把创建文件系统时的元数据映射保存到小文件内, 否则磁盘格式化过程将会失败, 导致无法进行数据读写.

NTFS 是 Windows 操作系统下常用的文件系统, 我们首先在该文件系统下进行了地址映射测试. NTFS 文件系统格式由四部分组成, 通过获取文件系统创建时的 * off 和 * len, 即可以计算出 2 TB 磁盘的文件系统元数据写入的位置和大小, 这样就可以把这四部分映射到实际物理地址, 完成系统格式化过程. 容量为 2 TB 的虚拟磁盘地址映射方案如图 2 所示.

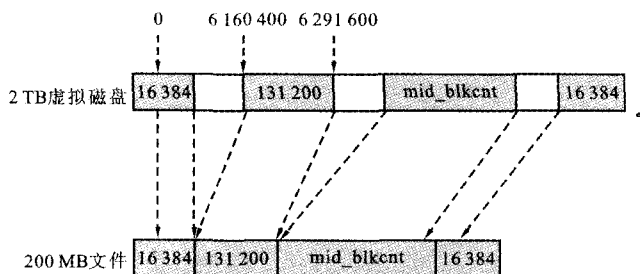


图 2 NTFS 地址映射方案

在地址映射过程中, 对除 mid_blkcnt 部分之外的三部分进行大小固定的地址映射, 文件剩余的容量都参与到 mid_blkcnt 部分的映射. 在对磁盘进行读写时, 如果返回 0 则表示成功, 对于无法建立实际地址映射的部分可以采用成功返回的方法来模拟对磁盘的读写测试.

3.3 动态地址映射

3.2 节中提出了一种应用于 NTFS 文件系统的地址映射方案, 然而 iSCSI Initiator 除了应用于 Windows 操作系统之外, 还可以应用于 Linux 操作系统. 由于不同文件系统的格式并不一样, 上述的地址映射方案就不具有普遍性. 我们借鉴精简配置中

按写分配的思想, 提出并实现了一种动态的地址映射机制^[4-5].

按写分配即先写先分配, 把虚拟磁盘划分为一定大小的块, 每次对一个新块进行写操作时就为其分配实际物理空间. 块在虚拟磁盘和实际磁盘之间的映射需要一个映射表来进行维护. 该映射表在启动 Target 服务时初始化在磁盘文件的最前端, 块的大小设置由程序员决定. 动态地址映射方案如图 3 所示.

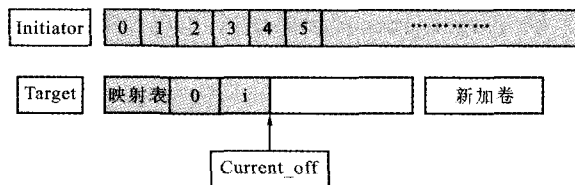


图 3 动态地址映射

映射表中的每一个表项应该包含以下信息: 虚拟块首地址、实际块首地址、下一块首地址和该块的可用标识. 比如当先写第 0 块时, 该块就会映射到映射表后; 写第 5 块时, 则会映射到实际磁盘的第二块. 这样地址映射一旦建立就可以完成对相应块的读写操作. 其中 Current_off 保存了当前实际磁盘已使用大小的偏移, 当 Current_off 值大于等于实际磁盘的大小时 (磁盘实际大小在初始化时可以获得, 并保存在相应 lun 结构体中), 同 3.2 节中采用成功返回 0 的方法保证磁盘读写测试的顺利进行.

3.4 动态扩容

在 3.3 节的基础上我们又设计了一种动态扩容方案. 随着按写分配资源的增加, Current_off 值将增大到磁盘的实际容量值, 而 Initiator 端由于容量虚拟仍会有较大的寻址空间, 此时可以打开新的磁盘文件逻辑上添加到 Target 磁盘的后边 (如图 3 中新加卷), 使得 Current_off 指针指向新的逻辑盘, 继续动态地址映射过程, 模型设计如图 4 所示.

图 4 中我们采用了常见磁盘阵列的磁盘组织结构. DMF 为动态扩容中的卷管理模块, 由于一个虚拟磁盘可能需要对应几个逻辑卷, 因此必须对卷之间的对应信息进行管理. 当可用卷不足时, 需要调用磁盘管理接口生成新的虚拟卷, 添加到 Target 卷的后边. 由于不同的磁盘阵列接口也不同, 我们只是在虚拟卷 (Virtual Volume) 层进行了模拟测试. 当前卷的容量不足时, 即 Current_off 值指向磁盘末端时, 我们使用 filp_open() 函数打开新的虚拟卷, 并将 Current_off 指针更新到新添加卷头部, 继续执行磁盘读写.

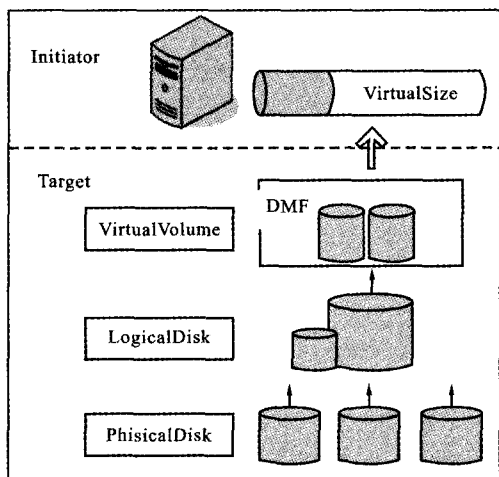


图4 动态扩容模型

4 实验

4.1 测试结果

IOmeter 工具是磁盘 IO 性能测试软件,可以测试磁盘的每秒最大 IO 次数 IOPS 和最大数据传输率 MBPS,测试时可以设置读写块的大小,读写比例和随机度等参数。

通常测试最大 IOPS 可采用 512 B, 100% Read, 0% Random 模式;测试最大数据传输率可采用 64 kB, 100% Read, 0% Random 或者更大的块。在本次实验中使用 100% Read 和 0% Random 模式,读写块大小分别取 512 B、1 kB、2 kB、4 kB、8 kB、16 kB、32 kB 和 64 kB,分别测试 IOPS 和 MBPS 变化趋势,实验结果如图 5、图 6 所示。

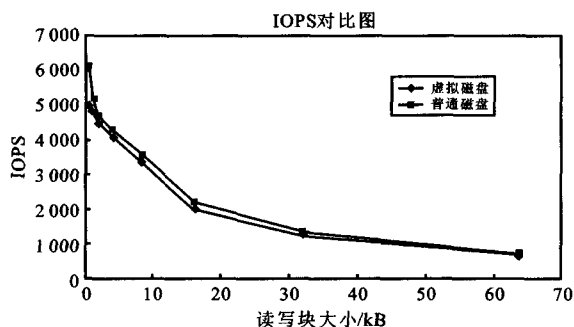


图5 IOPS对比图

4.2 结果分析

在 IOPS 对比图中,可以看到随着块大小的增加 IOPS 值相差越小,因为每一次块的读写都要对应一次地址映射,块越小,每秒的读写次数就越多,地址映射的时间开销所占比例就越大,因此块大小为 512 B 时两个曲线相差最大。

在 MBPS 对比图中,块越小二者相差越少,因为 MBPS 的计算公式表示如下:

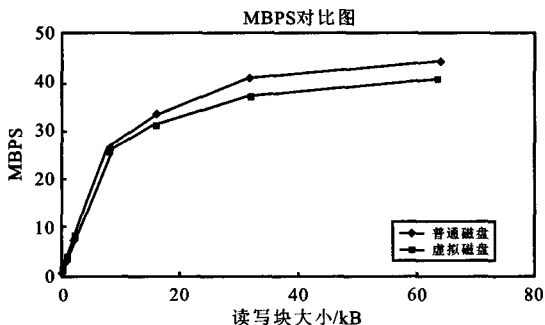


图6 MBPS对比图

$$\text{MBPS} = \text{IOPS} \times \text{块大小}$$

可知,MBPS 受两方面的影响。由 IOPS 对比分析可知,块尺寸较小时,IOPS 虽然相差较大,可二者乘积相差却较小,此时块大小影响因素更大;当块尺寸增大时,虽然 IOPS 相差变小,但乘积相差却增大,可知此过程块的大小占了主要因素。当块大小进一步增大时,IOPS 相差越来越小,此时两者乘积的大小无法定量的给出,从 MBPS 对比图来看,两者性能在达到最大之后随着块大小的增加,两条曲线有靠近的趋势。

5 结束语

本文通过分析 iSCSI 的通信机制和地址映射原理,实现了从小文件到大容量磁盘的模拟仿真,解决了海量存储系统硬件环境不足的难题。通过分析文件系统的格式,实现了更具通用性的动态地址映射方案,使得这种容量模拟仿真的方法更具有实用性,在此基础上进一步提出了基于 iSCSI 的动态扩容机制,并对这种扩容机制进行了模拟实现和测试。最后,对仿真磁盘和普通磁盘的性能进行了对比分析,证明了该磁盘仿真方案的正确性和合理性,尽管在性能方面与实际磁盘略有不同,但在相同的仿真硬件环境下仍可对系统的功能和性能趋势方面做出正确的评判,起到了非常重要的作用。

参考文献:

- [1] 鞠大鹏,刘川意,汪东升,等. 新型 IP 网络存储器性能比较[J]. 微电子学与计算机, 2008, 25(9): 190-192.
- [2] 丁双喜,张盛兵,贾宝峰. 双通道流水线 Flash 存储系统的设计[J]. 微电子学与计算机, 2008, 25(8): 224-228.
- [3] 姚玉坤,黄伟,杜宁. iSCSI 协议的仿真实现[J]. 计算机工程, 2007, 33(8): 241-243.
- [4] 朱海东,王新社. 基于扇区映射的 NOR flash 磁盘模拟方法[J]. 计算机工程与设计, 2007, 28(24): 5899-5902.

(下转第 66 页)

5 结束语

文中提出了一种电压跟随能力强、低功耗、调节范围大的电压控制 CMOS 电调谐第二代电流传输器(ECCII),通过引入对称的 CMOS 电流舵电路,保证了电流传输精度,电流增益连续可调,调节因子 $2 \leq K \leq 2$,同时避免了电流镜的过多使用,减小了电流损耗.采用电压调控方式,增大了电流输入范围.将电流输入端和电流加减电路隔离,保证了精确的电压跟随能力.采用 TSMC 0.35 μm 工艺参数,在 $\pm 1.5\text{ V}$ 电源供电的条件下对电路进行了 Hspice 模拟, V_Y/V_X 和 I_Z/I_X 的 -3 dB 带宽分别为 83.5 MHz 和 136 MHz ,功耗为 1.7515 mW . X 端的输入电阻不高于 22.8529 ,电流输出端 Z 端的输出阻抗不低于 $197.8336\text{ k}\Omega$, Y 端的阻抗为 $1\text{e}20\Omega$.通过验证,该电路可用于可调谐连续时间电流模式滤波器的设计.

参考文献:

- [1] Shahram Minaei, Onur Korhan Sayin, Hakan Kuntman. A new CMOS electronically tunable current conveyer and its application to current-mode filters[J]. Kuntman IEEE Trans Circuits Syst, 2006, 53(7): 1448-1456.
- [2] Abuelmaattim T, Bentrucia A, Alshahrani SM. A novel mixed mode current conveyor based filter[J]. International Journal of Electronics, 2004, 91(3): 191-197.

- [3] Shahra Minaei, Onur Korhan Sayin, Hakan Kuntman. A new CMOS electronically tunable current conveyer and its application to current-mode filters[J]. IEEE Transactions on Circuits and Systems, 2006, 53(7): 1448-1457.
- [4] Surakamponporn W, Kumwachara K. CMOS based electronically tunable current conveyer[J]. Electronic Letters, 1992, 28(14): 1316-1317.
- [5] 宋树详, 严国萍. 一种新型 CMOS 电调谐第二代电流传输器(ECCII)[J]. 半导体技术, 2007, 32(12): 1077-1080.
- [6] Jaime RamirezAngulo, Garimella Sri Raga Sudha, Garvajal R G, et al. New gain programmable current mirrors based on current steering[J]. IEEE Trans Circuits Syst, 2006, 2(49): 162-165.
- [7] 彭建华, 陈向东, 杜广涛. 一种轨对轨输入的低压低功耗 CMOS 电流传输器[J]. 微电子学与计算机, 2008, 25(11): 72-76.

作者简介:

李沁莲 女, (1985—), 硕士研究生. 研究方向为低压低功耗 CMOS 模拟集成电路设计.

陈向东 男, (1967—), 教授, 博导. 研究方向为微电子技术、信息获取技术及新型传感器.

王丽娜 女, (1986—), 硕士研究生. 研究方向为低压低功耗 CMOS 模拟集成电路设计.

梁建 男, (1985—), 硕士研究生. 研究方向为低压低功耗 CMOS 模拟集成电路设计.

(上接第 61 页)

- [5] IETF RFC 3720, Internet small compute system interface (iSCSI)[S]. IETF, 2003.

作者简介:

田龙 男, (1984—), 硕士研究生. 研究方向为软件工程与

网络软件.

张延园 男, (1954—), 教授. 研究方向为网络存储、软件工程.

张晓 男, (1978—), 副教授. 研究方向为网络存储.