

iSCSI 性能优化方法综述

谭志虎 万继光

(华中科技大学计算机学院信息存储系统教育部重点实验室 武汉 430074)

(武汉光电国家实验室 武汉 430074)

摘要 iSCSI是利用TCP/IP网络进行SCSI数据传输的协议,与传统存储访问协议相比,其不受限于连接距离,连接方式和拓扑结构灵活,硬件成本和管理成本大大降低,但由于TCP/IP协议栈以及数据拷贝等开销,iSCSI数据传输性能存在一定的瓶颈。详细分析和探讨了目前优化iSCSI传输性能的几种主要方法,并探讨了iSCSI协议未来发展趋势。

关键词 iSCSI,性能优化

Review of iSCSI Performance Improvement Approaches

TAN Zhi-hu WAN Ji-guang

(Key Laboratory of Data Storage System, Ministry of Education, Huazhong University of Science and Technology, Wuhan 430074, China)

(Wuhan National Laboratory for Optoelectronics, Wuhan 430074, China)

Abstract iSCSI protocol is a transfer for SCSI over TCP/IP network. Compared with traditional storage access protocol, it has not limitation in the distance, and provides flexible connection method and topology, and significantly reduces the hardware cost and management cost. However, due to the cost on TCP/IP protocol stacks and data copy, some bottleneck exists on the performance of iSCSI data transfer. This paper illustrated and discussed currently several methods to optimize the iSCSI data transfer performance, and also discussed the future trend of iSCSI.

Keywords iSCSI, Performance improvement

1 引言

为解决传统并行SCSI协议连接距离及连接设备数有限的问题, SAM-2协议定义了SCSI传输层体系结构,该协议包括串行SCSI(SAS)、光纤通道(FCP)、iSCSI等协议。其中iSCSI是SCSI到TCP/IP协议的映射,是利用TCP/IP网络进行SCSI数据传输的协议,其将标准的SCSI CDB数据包封装成IP数据包,在IP网络上进行传输,到达目的地后,再恢复成封装前的SCSI命令,从而实现SCSI命令在IP网络上的直接透明传输。iSCSI目标器与启动器之间的通讯是通过一个或多个TCP连接完成的, TCP连接被用来传输控制信息、SCSI命令和参数以及封装在iSCSI协议数据单元(PDU)中的数据。

与FCP协议相比iSCSI协议具有以下优势:1. 成本低廉,利用iSCSI技术构建IP-SAN系统不需要昂贵的专用光纤存储网络及光纤交换设备,采用传统的以太网即可构建,相比FC-SAN系统具有较强的价格优势;2. 管理成本降低,采用iSCSI构建的IP-SAN系统甚至可以利用传统的SNMP协议进行管理,大大降低了管理复杂度;3. 协议通用,iSCSI协议可以利用传统IPSec协议进行安全控制,还可利用SLP(Service Location Protocol)协议进行设备发现;4. 无限扩展连接距离,尤其适合远距离数据备份与恢复,相比FCP有限的传输距离而言,iSCSI连接距离是无限可扩展的。

2 iSCSI 体系结构

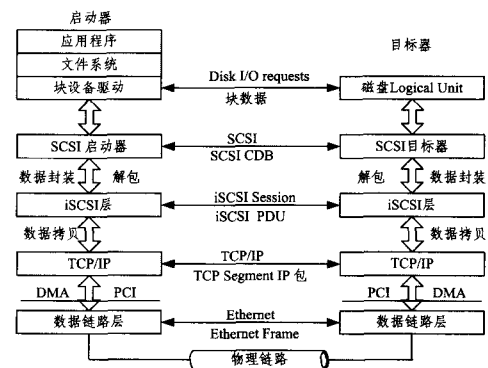


图1 iSCSI 体系结构模型

图1为iSCSI体系结构模型图,与传统SCSI存储设备访问一样,iSCSI协议中发起访问的一方称为启动器,接受访问的存储设备端称为目标器。应用程序需要访问目标器的磁盘逻辑单元时,由文件系统及块设备驱动将磁盘I/O请求转换成一条或多条SCSI命令(SCSI CDB),iSCSI层接收到SCSI层传递过来的SCSI CDB后,将SCSI CDB封装成iSCSI协议数据单元PDU,然后通过TCP/IP层将PDU发送到以太网,各层依次对上层传递过来的数据包进行各自的封装,目标器端则进行着与启动器端完全相反的操作,当其从TCP/IP

到稿日期:2008-02-22 本文受国家自然科学基金(60603075)与国家“九七三”重点基础研究发展计划(2004CB318203)资助。

谭志虎(1974—),男,博士,研究方向为计算机网络存储,E-mail: stan@mail. hust. edu. cn;万继光(1972—),男,博士,研究方向为计算机网络存储。

层接收到 iSCSI PDU 后将其进行解包后发送给 SCSI 目标器层直至物理磁盘设备,完成数据访问后目标器将通过 TCP/IP 层返回响应数据或状态信息给启动器。启动器与目标器双方 iSCSI 层之间的通信称为会话(Session),一次会话包括一个或者多个 TCP 连接,会话包括登录以及实际 SCSI 处理两个阶段,登录的主要作用是确认访问权限和以及后续实际 SCSI 命令数据传输参数。

3 iSCSI 性能瓶颈

iSCSI 协议建立在传统的 TCP/IP 协议之上,在进行实际数据传输时,其传输系统性能受限于 TCP/IP 协议栈负载及以太网最大带宽,另外 iSCSI 协议层也会额外增加一些负载开销,在实际应用中,iSCSI 数据传输性能仍然存在瓶颈。

以启动器端写操作为例,一次完整的 iSCSI 协议数据传输包括数据封装、数据拷贝、数据传输 3 个步骤:①数据封装,iSCSI 协议层首先将接收到的 SCSI 命令封装成 iSCSI 协议层的 PDU 数据包,除了一般的 PDU 协议头的封装外,此过程还需要对 PDU 数据包进行 CRC 校验计算,以便目标器接收到数据后进行正确性检测;②数据拷贝,封装完毕后的 PDU 数据包以数据拷贝的方式传递给操作系统的 TCP/IP 协议层缓冲区,由其发送给数据链路层进行网络传输,一个 PDU 数据有可能对应多个 TCP 连接;③数据传输,TCP/IP 层接收到数据包后再通过 PCI DMA 的方式将数据传递给网卡经由物理链路传输到目标器。读操作过程操作顺序正好相反,目标器端操作也与启动器端类似。

数据封装过程必须进行 CRC 循环冗余校验计算,校验生成与检测属于计算密集型操作,需要消耗大量 CPU 资源,会严重影响 iSCSI 系统的性能;数据拷贝过程也会占用启动器或目标器端大量的 CPU 资源及内存带宽,从而影响系统性能;Abhijeet^[1]给出了在 1.7GHz Pentium 处理器,400MHz 前端总线,单通道 DDR266 环境下具体的一个性能测试,该测试表明 CRC32 是 iSCSI 协议处理中最耗时的部分,平均 6.5 个时钟周期才能处理一个字节,而内存拷贝则是每 2.2 个时钟周期处理一个字节。数据传输过程利用 TCP/IP 协议进行网络传输,其必然受限于 TCP/IP 协议栈本身的开销负载以及实际网络带宽,相关试验表明每赫兹主频仅能传递一个 bit,故利用 TCP/IP 进行数据传输也会严重占用系统 CPU 资源。

4 iSCSI 性能优化

如何进一步优化现有 iSCSI 技术,解决其潜在性能瓶颈问题,是 iSCSI 研究领域的热点问题,根据 iSCSI 协议在数据传输、数据拷贝、数据校验 3 个方面潜在的性能瓶颈问题,目前对 iSCSI 协议进行性能优化的方法主要分为以下 3 种。

4.1 优化网络传输

iSCSI 协议利用通用的 TCP/IP 协议进行海量存储数据传输,传统 TCP/IP 协议用于在不可靠的数据链路上实现可靠的数据传输,其性能必然受限于 TCP/IP 本身的极限传输性能,基于通用 TCP/IP 的存储网络一个较重要的性能问题就是 TCP/IP 协议栈开销。目前已经出现了一些新型的网络技术用于改善 TCP/IP 协议栈开销:

1. 校验计算卸载。一些新型网卡可以分担过去由主机

CPU 完成的 CRC 校验和计算功能,可进一步释放主机 CPU 资源,降低 CPU 利用率。

2. TSO 技术。当数据包大于 TCP 包时,必须进行分拆,一些新型号网卡可以利用网卡实现过去由主机 TCP/IP 软件协议栈完成的 TCP 包分拆功能,也可缓解主机 CPU 压力。

3. JUMBO 巨型帧。采用大的 TCP/IP 数据包可有效优化网络传输性能,以太网最大传输单元 MTU 最大为 1500 字节,如果传输数据超过这个大小,则需要对数据进行分割并重新封装,分割及数据操作重新封装均会消耗 CPU 资源,为了降低该操作对 CPU 资源的影响,一些新型 GB 级以太网适配卡引入了 9000 字节的 JUMBO 巨型帧,巨型帧的引入可有效减少分割封装次数,从而优化系统性能。

文献[2]指出在实际的 iSCSI 传输中,58% 的 TCP/IP 数据包均小于 127 字节,为了避免小数据包请求对 TCP/IP 网络的占用,其在启动器端设置一个以日志文件系统进行组织的磁盘缓存区——iCACHE,其将启动器端多个小写请求组合成大写请求缓存起来,利用有效的 CACHE 调度算法使得大多数小数据包请求均可在启动器端完成,系统空闲时再将脏数据从缓冲区迁移到远端目标器存储设备上,可有效减少对 TCP/IP 网络的占用,进一步提升 iSCSI 性能。

4.2 避免数据拷贝

传统 SCSI 协议层直接与主机适配卡硬件进行数据交互,而 iSCSI 层建立在传统 TCP/IP 协议层之上,其 PDU 数据包必须通过内存拷贝的方式传递给 TCP/IP 协议栈缓冲区进行网络传输,内存拷贝涉及内存读写操作各一次,内存数据拷贝必须依赖于 CPU 执行具体指令完成,会消耗 CPU 资源并降低内存性能从而影响 iSCSI 系统性能。为了避免数据拷贝操作,目前主要有以下 3 种方法:

1 零拷贝技术

Ogawara^[3]利用虚拟内存重映射技术实现了 iSCSI 协议层与通用 TCP/IP 协议栈之间的零拷贝数据传输,该技术使得 iSCSI 层与 TCP/IP 层之间通过交换数据地址的方式进行数据交互,有效避免了数据拷贝,其试验结果表明采用该技术后系统 CPU 占用率从 39.4% 下降到 30.8%。但该方法必须修改操作系统 TCP/IP 协议栈内存管理以及虚拟内存 Cache 管理系统,不具有实用性,仅仅停留在实验室研究领域。

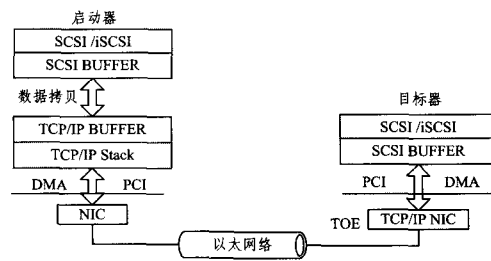


图2 TOE 硬件方案

2 TOE 硬件方案

TOE 方案将过去由操作系统完成的 TCP/IP 协议栈直接下移到网卡由硬件实现,如图 2 所示。该方案有效避免 TCP/IP 软件协议栈对系统 CPU 资源的占用,而 iSCSI 层与 TCP/IP 层之间的数据拷贝也由内存数据拷贝变成 PCI DMA 操作,可进一步优化系统性能,但这种方案不能充分利用 CPU 硬件性能逐渐提升带来的益处,对一些特殊协议的处理

可能还会成为系统瓶颈^[4]。

3 拷贝和校验交替

对于写操作,iSCSI层PDU数据经过CRC校验计算后以多个TCP包的形式发送给TCP/IP协议栈,读操作正好相反,完整的PDU数据包经多个TCP包拼装后再进行CRC校验检测,数据拷贝操作与完整PDU数据包的校验计算操作是串行进行的。Abhijeet^[1]中提出了PDU数据校验计算与数据拷贝交替进行的方式,数据拷贝与数据校验均以字节为单位,写操作中,每完成一个字节的校验计算,即向TCP/IP协议栈拷贝发送一个数据,而读操作中,TCP/IP协议栈层每拷贝一个字节到iSCSI层,即进行一个字节流的CRC校验检测,校验计算是计算密集型操作,而数据拷贝是内存访问密集型操作,两种操作交替进行可充分利用CPU的处理能力,另外两个操作涉及对同一数据区域的操作,也可有效发挥Cache的效率,试验结果表明采用这种方法CPU周期数减少32%。但这种方法必须修改iSCSI层与TCP/IP协议栈之间的接口,使得二者均能适应数据拷贝与数据校验的交替操作。

4.3 数据校验的优化

PDU数据包中包含CRC32校验字段,用于保证PDU数据传输的可靠性,通常是用反馈移位寄存器实现,其硬件电路设计较为简单,但在iSCSI PDU协议处理中这部分计算必须用软件长除法完成,待校验的数据必须与生成多项式进行逐位的模二除法(异或运算),每运算一次校验数据右移一位,直至所有数据计算完毕才能得到最终的CRC校验数据。利用长除法计算CRC校验值每进行一次异或运算仅能处理一位,CRC校验数据生成或检测均是计算密集型操作,会消耗大量的CPU资源,因此如何优化CRC32数据校验性能成为优化iSCSI性能的关键,目前主要有以下3种方法优化CRC校验性能。

1. 优化CRC校验算法。Abhijeet^[1]引用了一种特殊的SB8(Slicing-8)CRC校验计算方法,其改变了传统逐位除法的模式,采用查表方式进行运算,将任意8位数据左移32位后的CRC32校验值事先计算好存放在查找表中,其运算步骤如下:①32位校验值C初始化;②将校验值C的首字节与待校验数据首字节作异或运算;③将运算结果作偏移值在查找表中查找结果,查找到的结果与左移8位后的校验值进行异或后生成新的校验值;④跳转至步骤②处理待校验值字节流第二个字节,直至所有数据处理完毕。传统长除法每次运算仅能处理一个比特,而这种新SB8校验计算方法每次运算可以处理一个字节,处理一个字节数据仅需要两次异或运算和一次查表操作,由于查找表体积较小,经过一段时间访问后查找表内容将会被调度到CPU二级缓存,可有效优化查表速度。采用这种优化的算法将会比传统校验算法快3倍左右。

2. iSCSI HBA 硬件方案。为了进一步消除CRC校验运算对系统CPU资源的影响,工业界将iSCSI协议以及TCP/IP协议全部集成到网卡上构成iSCSI HBA适配卡,如图3所示。这种方式中iSCSI HBA适配卡对主机来说就是一块SCSI卡,既不存在数据拷贝问题,也不存在数据校验问题。但

采用硬件方案虽然能进一步释放CPU资源,但iSCSI HBA硬件处理能力有限制,无法分享CPU日益提升的性能^[5,6]。

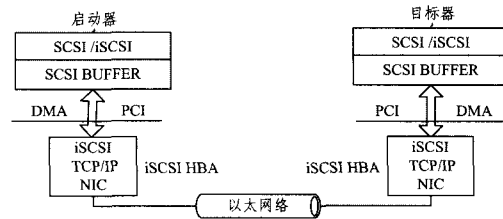


图3 iSCSI HBA 硬件方案

3. 直接关闭校验。CRC校验主要保证PDU数据在传输中的可靠性,避免数据被篡改,如果不强调安全性,CRC校验意义不大,所以在有些对性能要求较高而对数据安全性要求不高的环境中,可以通过关闭CRC校验的方法优化iSCSI性能,数据的正确性依靠TCP/IP层的CRC校验数据来保证。

结束语 利用iSCSI技术构建的基于IP的存储网络系统,无需昂贵的专用存储网络,易于管理维护,可有效减少用户投资,其具有良好的扩展性,方便进行异地容灾,随着10G以太网的出现,iSCSI数据传输性能瓶颈问题也有了较大的改观,Abhijeet^[1]测试结果表明iSCSI传输性能已经可以高达3.6Gbps。对iSCSI存储系统的优化,主要集中在对iSCSI数据校验、数据拷贝、数据传输3个方面进行改善,随着iSCSI技术日益成熟,采用iSCSI技术构建的IP-SAN系统将拥有广阔的应用前景。

参考文献

- [1] Joglekar A, Kounavis M E, Berry F L. A Scalable and High Performance Software iSCSI Implementation // 4th USENIX Conference on File and Storage Technologies; 267-280
- [2] He Xubin, Yang Qing, Zhang Ming. A caching strategy to improve iSCSI performance // Local Computer Networks, 2002. Proceedings. LCN 2002. 27th Annual IEEE Conference. Nov. 2002; 278-285
- [3] Tomonori F, Masanori O. Performance of optimized software implementation of the iSCSI protocol // Proceedings of the international workshop on Storage network architecture and parallel I/Os. New Orleans, Louisiana, September 2003; 1-10
- [4] Sarkar P, Uttamchandani S, Voruganti K. Storage Over IP: When Does Hardware Support Help? // Proceedings of the 2nd USENIX Conference on File and Storage Technologies, San Francisco, CA, March 2003
- [5] Aiken S, Grunwald D, Plszkun A R, et al. A Performance Analysis of the iSCSI Protocol // Proceedings of the 20th IEEE/11th the NASA Goddard Conference on Mass Storage Systems and Technologies (MMS'03)
- [6] Xinidis D, Bilas A, Flouris M D. Performance Evaluation of Commodity iSCSI-based Storage Systems // Proceedings of the 22nd IEEE/13th NASA Goddard Conference on Mass Storage Systems and Technologies (MSST'05)