

无线网络中 iSCSI 性能的研究

李建设, 文 鸿, 董 刚

(湖南工业大学 计算机与通信学院, 湖南 株洲 412008)

摘 要: 在介绍和分析标准 iSCSI 协议体系结构的基础上, 给出了在 iSCSI 会话中通过建立多个虚拟 TCP 连接且在无线局域网中的 iSCSI 层采用并行工作机制的优化方法, 并对其性能进行了理论分析和实验测试. 结果表明, 与标准 iSCSI 相比, 在无线网络中使用多连接 TCP 能有效的改善 iSCSI 协议的存储性能.

关键词: 多连接; 并行机制; iSCSI; 无线网络

中图分类号: TP393.1

文献标识码: A

文章编号: 1000-7180(2011)07-0156-03

Study of iSCSI Performance in Wireless Network

LI Jian-she, WEN Hong, DONG Gang

(School of Computer and Communication, Hunan University of Technology, Zhuzhou 412008, China)

Abstract: This paper introduces and analyzes the normal iSCSI protocol architecture, presents the optimization method of multiple virtual TCP connections in an iSCSI session and parallel working mechanism in iSCSI layer over wireless LAN, the performance of the multiple connections is analyzed theoretically and tested. The results have shown to reach the conclusion that the multiple connections can improve a storage performance of the iSCSI protocol in wireless network.

Key words: multiple connections; parallel mechanism; iSCSI; wireless network

1 引言

为适应网络环境中大量信息的存储和管理需求而建立的存储区域网(SAN),使基于以太网的网络存储迅速的流行起来,为此,工业上已开发出 FCIP、mFCP、iFCP、iSCSI 和 iSNS 等各种不同的技术,其中,最主要的两个核心协议是 FCIP 和 iSCSI. FCIP 用于构建基于 IP 网络的 FC-SANs,而 iSCSI 则是通过整合存储协议 SCSI 和网络协议 TCP/IP,实现了存储与 TCP/IP 网络的无缝融合. 与 FCIP 相比, iSCSI 不仅使用成本低,而且还具有很好的可扩展性以及远程备份和灾难恢复实现简单等优点.

文中在分析标准 iSCSI 协议体系结构的基础上,给出在 iSCSI 会话中建立多个虚拟 TCP 连接,同时在 iSCSI 层中采用并行工作机制的优化方法,以提高 iSCSI 在无线网络中的存储性能,并对其性

能进行了理论分析. 最后,对单连接的标准 iSCSI 协议和多连接 iSCSI 在无线局域网中的吞吐量、响应时间和 CPU 效率等性能进行了测试和对比研究.

2 标准 iSCSI 体系结构

构建在 TCP/IP 上的 iSCSI 协议体系结构,如图 1 所示. 启动器(Initiator)和目标器(Target)之间通过 iSCSI 协议建立的会话进行数据和命令的交换. 启动器中,若应用程序要访问目标器中存放的文件,就会发出读文件的请求,并将请求传给其下层的文件系统. 由文件系统将文件请求转化成块请求并传送到 SCSI 层,SCSI 层会将其转换为 SCSI 命令(SCSI CDB)后送到 iSCSI 层, iSCSI 层接到后,将其封装成 iSCSI 协议数据单元并通过 TCP/IP 层将其发送到网络上. 目标器从 TCP/IP 层接收到启动器发来的读文件请求 iSCSI PDUs 后, iSCSI 驱动程序

收稿日期: 2010-10-18; **修回日期:** 2010-11-06

基金项目: 国家自然科学基金资助项目(60773110);湖南省教育厅资助科研项目(09C328)

首先对其进行解封装,然后将 SCSI 命令映射到存储设备. 目标器将获得的应答数据或状态通过 TCP/IP 层发回到启动器. SCSI 命令的执行可以概括为发送命令、获取数据以及数据或状态反馈三个阶段. 由图 1 可知,在一次 I/O 访问期间,如果是对目标器主存(即 RAM I/O)的读写操作,要进行一次数据复制和一次 DMA 操作,而对磁盘读写(即 Disk I/O),则需进行两次数据复制和一次 DMA 操作.

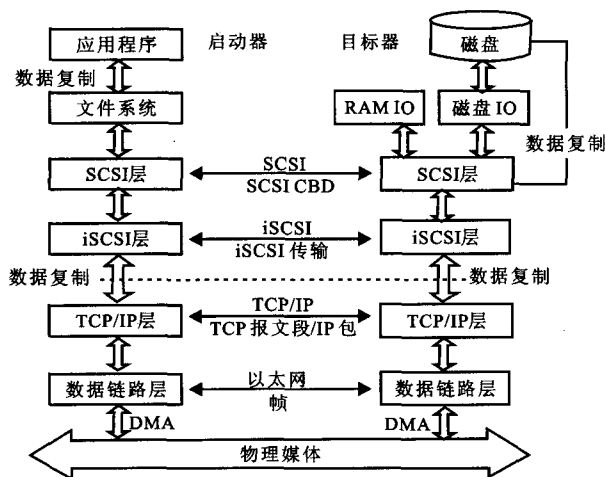


图1 标准 iSCSI 体系结构

建立在 TCP/IP 协议之上的 iSCSI,利用 TCP 对 SCSI CDBs 进行封装来实现数据传输,其性能受限于 TCP/IP 协议栈负载及最大网络带宽. 另外,标准 iSCSI 的执行基于单个 TCP 连接,TCP 窗口大小以及在发送数据包前,要收到已发送数据包的 TCP 确认的限制,使单个 TCP 连接无法充分利用整个网络带宽. 特别是在低速且不可靠的无线环境下,其性能还将面临数据包丢失和长时延所造成的更加严重的影响.

3 多连接 iSCSI 的并行优化

由以上分析可知 iSCSI 协议性能主要受限于 TCP 性能. 由于 iSCSI 协议支持在一个会话中的多个虚拟 TCP 连接^[1-2]. 因此可使用多个虚拟 TCP 连接的并行工作机制,并在 iSCSI 层中采用并行工作机制进行数据传输^[3-4],从而达到改善 iSCSI 性能的目的.

3.1 工作原理

建立在一个物理连接上的多个虚拟 TCP 连接中,有一半的连接用于从启动器发送 SCSI 请求到目标器,另一半则用来从目标器发送响应到启动器. 一对分别位于启动器和目标器的发送线程和接收线程负责一次数据和命令的传输,其工作原理如图 2 所示.

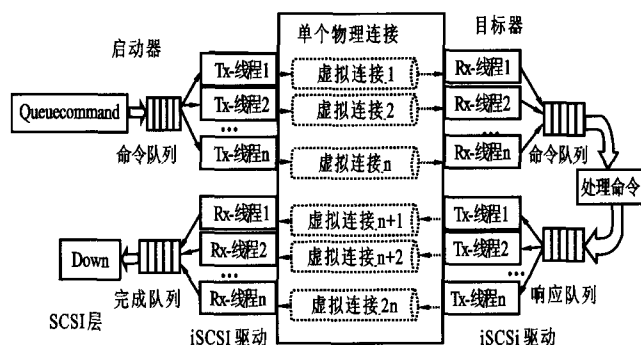


图2 多并行 TCP 连接 iSCSI 体系结构

在标准 iSCSI 体系结构中 SCSI 层由 SCSI 的上子层、中间子层以及下子层驱动器构成. SCSI 上子层接受来自 SCSI 上层,即文件系统层的请求并将其转换成 SCSI 请求交给 SCSI 中间层,同时还负责将返回的状态信息通知文件系统层. SCSI 中间子层是 SCSI 上子层和下子层的公共服务层,它提供了很多可供其上子层和下子层使用的函数,因而它是 SCSI 上下子层进行通信的标准接口. 下子层的由一组驱动程序构成,负责处理各种不同类型的 SCSI 适配器. 以读操作为例说明整个通信过程. SCSI 中间子层的两个主要函数分别是 queuecommand() 和 done(), 其中 queuecommand() 用于向 iSCSI 驱动程序发出 SCSI 命令,而 done() 则用于通知 SCSI 中间子层所收到的 SCSI 命令已由 iSCSI 驱动程序完成. iSCSI 启动器中 SCSI 下子层驱动程序从 SCSI 中间子层接到读命令,并将这些封装在 iSCSI 请求协议数据单元中命令在启动器中的命令队列中排队.

在启动器中,虚拟连接所对应的 Tx 线程从命令队列中取出协议数据单元发送到目标器,目标器驱动程序通过其连接所对应的 Rx 线程接收协议数据单元,然后对 iSCSI 请求协议数据单元进行解封装,并将 SCSI 命令映射到存储设备(RAM 或 disk). 从存储设备获得数据,并形成 iSCSI 响应协议数据单元(包括数据和状态)之后,目标器驱动程序将这些 iSCSI 响应协议数据单元在响应队列中排队. 目标器的 Tx 线程通过对应连接,将这些协议数据单元发送出去. 启动器驱动程序通过其连接所对应的 Rx 线程接收,然后将其放到完成队列中,最后调用 done() 函数完成 SCSI 交换.

在使用多虚拟 TCP 连接时,多个命令和数据同时并行传输. 在这种情况下,启动器和目标器必须能够识别每个收到的 PDU. iSCSI PDU 的基本头结构(Basic Header Structure, BHS)中有一个 4 字节的启动器任务标识(Initiator Task Tag, ITT)字段,用

于为每个 iSCSI 任务分配一个唯一标识,即在每个 iSCSI PDU 中记录 iSCSI 任务的顺序编号. 由于 iSCSI 的每个任务是由一个 iSCSI 命令发起的,且该命令可能有许多与其相关 SCSI 活动,所以同一 iSCSI 命令启动的任务所对应的相关数据和状态 PDUs 都使用相同的 ITT,即 ITT 值在当前会话的上下文中是唯一的. 这样,在多连接中启动器和目标器就能够很容易地从等待队列中找到每个任务各自的数据和状态 PDU 了.

3.2 理论分析

为从理论上分析多虚拟连接 iSCSI 的性能,分别建立标准 iSCSI 的单连接队列模型 M/M/1 和多虚拟连接 iSCSI 的多连接队列模型 M/M/m.

在 M/M/1 模型中,假设包的平均到达率是 λ , 服务时间服从指数分布,且其平均值为 $1/\mu_1$. 对于 M/M/m 模型,假设通信链路上有 m 个独立的服从泊松分布的通信流,包总的平均到达率为 λ . 链路被划分为的 m 个独立的通道,每个通道分配一个通信流. 如果一个通信流中没有包要传输,那么,它对应的通道可用于传送另一个通信流的包. 在每个通道中,服务时间服从指数规律,其平均值为 $1/\mu_m$.

单连接的平均时延可表示为

$$T_1 = 1/(\mu_1 - \lambda) \quad (1)$$

多连接的平均时延可表示为

$$T_m = 1/\mu_m + P_Q/(m\mu_m - \lambda) \quad (2)$$

多连接的排队概率 P_Q 表示为

$$P_Q = p_0(m\rho)^m/[m!(1-\rho)] \quad (3)$$

其中,服务强度 $\rho = \lambda/\mu_m$. 单个物理连接中的多虚拟连接,其 μ_m 的值依赖于工作负载的大小和连接数 m . 当连接数 m 为 2,系统工作在 $\rho \ll 1$ 的轻负载状态时,由方程(2)知, $P_Q \approx 0$ 且

$$T_2 \approx 1/\mu_2 \quad (4)$$

由此可知,在轻负载情况下,由于单个 TCP 连接不能充分的使用物理连接,多虚拟连接不需竞争单个物理连接且服务速率 μ_2 约等于 μ_1 . 从式(1)和(4)可看出,在轻负载状态下,M/M/2 模型中的包时延比 M/M/1 模型要小的多.

由方程(2),当系统工作在 ρ 接近 1 的重负载状态时, $P_Q \approx 1$, $1/\mu_2 \ll 1/(2\mu_2 - \lambda)$ 且

$$T_2 \approx 1/(2\mu_2 - \lambda) \quad (5)$$

对于重负载情况下,由于物理连接几乎都被单个 TCP 连接所占用,多虚拟连接通常需要竞争来使用一个物理连接且服务速率 μ_2 要比 μ_1 低很多,但比

$\mu_1/2$ 要大一些. 从方程(1)和(5)可以看出,在重负载情况下,M/M/2 模型中的包时延比 M/M/1 模型要稍小些.

在 iSCSI 中,小 I/O 请求块为轻负载,而大 I/O 请求块为是重负载. 通过上述分析可知,对于包传输,平均来说,多 TCP 连接的 iSCSI 比单连接的标准 iSCSI 的时延更小,也就使其在吞吐量等方面性能更好.

4 性能测试

测试平台是由运行 iSCSI 协议,使用 802.11b 进行通信的启动器和目标器构建的网络.

在无线网络环境下,多跳信道和不稳定信号强度会延长数据传输,使网络时延更长. 测试结果表明,多虚拟 TCP 连接使 iSCSI 在长时延的无线网络中获得了较高的吞吐量.

5 结束语

文中给出了多虚拟 TCP 连接的 iSCSI,并利用排队论对其性能进行了理论分析,同时还对其主要的存储性能进行了测试. 结果表明,与单个连接的标准 iSCSI 相比,使用多虚拟 TCP 连接的能有效地改善 iSCSI 协议在低带宽、不可靠、长时延的无线网络中的存储性能.

参考文献:

- [1] Julian Satran, et al, iSCSI (Internet SCSI) Specification, Draft [EB/OL]. [2010-03-13]. <http://www.ietf.org/internet-drafts/draft-ietf-ins-iscsi-Internet> 20. ht. 2003.
- [2] Meth K Z, Satran J. Features of the iSCSI protocol[J]. Communications Magazine, IEEE, 0 2003, 41(8): 72-75.
- [3] Xiong H, Kanagavelu R, Zhu Y, et al. An iSCSI design and implementation [C]// Proceedings of the 12th NASA Goddard & 21st IEEE Conference on Mass Storage Systems and Technologies (MSST2004). Maryland, USA; IEEE, 2004: 101-107.
- [4] Gao Yan, Zhu Yaolong, Xiong Hui, et al. An iSCSI design over wireless network [C]// Proceedings of 12th IEEE International Conference on Networks. Singapore; IEEE, 2004: 60-65.

作者简介:

李建设 男, (1966-), 硕士, 副教授. 研究方向为网络存储、无线网络与移动计算 (CCF 会员: E200014513M).