doi: 10.7690/bgzdh.2015.10.015

# 一种支持多协议的网络存储系统

李　丽，吴登勇，刘维霞
（山东超越数控电子有限公司研发中心，济南 250100）

**摘要：**针对目前以太网存储系统存在性能低和可扩展性差的问题，提出一种支持多协议的网络存储系统。从网络存储系统整体性能角度论证该系统硬件设计原则，给出多协议存储系统软件架构并进行性能测试。结果表明：该系统支持以太网、光纤和 ib_srp 网络存储协议，LIO 模式性能要优于 SCST 模式，是多协议网络存储系统的首选。

**关键词：**网络存储系统；多协议；光纤；infiniband
**中图分类号：**TP333　　**文献标志码：**A

# A Multi-protocol Network Storage System

Li Li, Wu Dengyong, Liu Weixia
(*Research & Development Center*, *Shandong Chaoyue Digital Control Electronics Co.*, *Jinan* 250100, *China*)

**Abstract:** For low performance and poor scalability of existing Ethernet-based storage systems, proposed a multi-protocol network storage systems. Demonstrating of system hardware design principles from the perspective of the overall performance of network storage systems and giving the multi-protocol storage system software architecture and performance testing. The results show that the system supports Ethernet, fiber channel and ib_srp network storage protocol. LIO has better performance than SCST mode and is the preferred option of multi-protocol network storage systems.

**Keywords:** network storage system; hybrid protocol; fiber channel; infiniband

## 0　引言

在存储系统发展的同时，处理器性能以更快的速度提升，多个处理器和多个虚拟机系统间的 I/O 共享在提高了资源利用率的同时，对存储系统的 I/O 能力施加了更大的压力[1]。

存储设备的访问速度提升远远落后于处理器和内存的提升。尽管存储设备中采用了各种缓存和预取[2]、并行执行[3-4]和各种 RAID 技术等方面的优化，大多数情况下的 I/O 瓶颈问题仍然归咎于存储设备中速度较慢的网络连接接口[5]。

目前，常用的网络连接口大多是以太网接口，这种接口的存储系统提供的主机接口相对单一，性能低、可扩展性差。为了解决这些问题，笔者提出了一种支持多协议的网络存储系统，该系统在存储设备前端预留了 3 个 PCIE 插槽，以支持以太网、光纤网和 infiniband 网多种连接方式，从而达到提高存储系统性能和可扩展性的功能。

## 1　相关技术

### 1.1　网络存储系统体系结构

存储系统体系结构先后经历了直接存储（direct attached storage，DAS）体系结构、网络附加存储（network attached storage，NAS）体系结构和存储局域网络（storage area network，SAN）体系结构 3 种主要类型的发展[6]。

DAS 是指将存储设备通过 SCSI 线缆或光纤通道直接连接到服务器上；NAS 是一种文件共享服务，拥有自己的文件系统，通过 NFS 或 CIFS 对外提供文件访问服务；SAN 是一种通过网络方式连接存储设备和应用服务器的存储构架，这个网络专用于主机和存储设备之间的访问。当有数据存取的需求时，数据可以通过存储区域网络在服务器和后台存储设备之间高速传输。图 1 对 3 种存储架构提供的服务层次进行了比较。

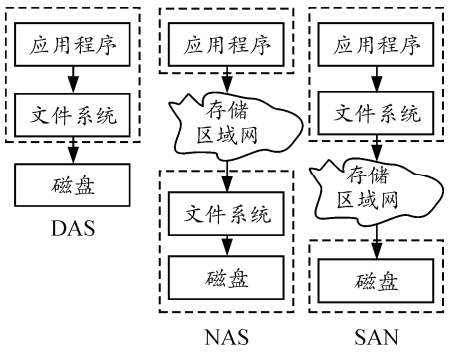


**图 1　3 种网络存储架构比较**

收稿日期：2015-05-17；修回日期：2015-06-30
作者简介：李　丽（1984—），女，山东人，硕士，工程师，从事存储软件、光纤存储、存储体系结构研究。

### 1.2　LIO 和 SCST 工作原理

当前 SAN 架构的存储系统实现主要依赖于 SCSI Target 软件，该软件是将服务器主机转为可透过不同网络协定远端存储设备的 SCSI 目标端存储装置。目前有 LIO 和 SCST（generic SCSI target subsystem for linux）2 种 SCSI Target，它们都是依据 SCSI 标准结构开发出来的，介于 Linux kernel 和后端存储控制器之间。图 2 为 2 种 target 的实现层次，从图中可见其核心模块实现于 Linux 存储结构中块设备的上层，用于接收 I/O 命令并将它们传递到 SCSI 的中间层。

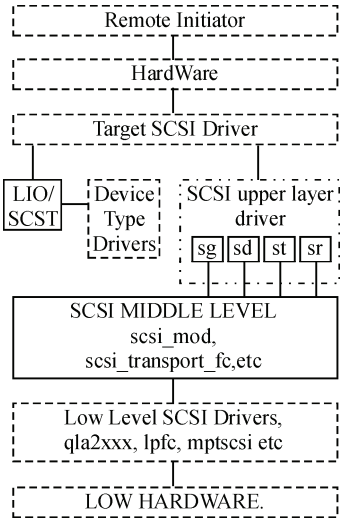


图 2　LIO/SCST 结构

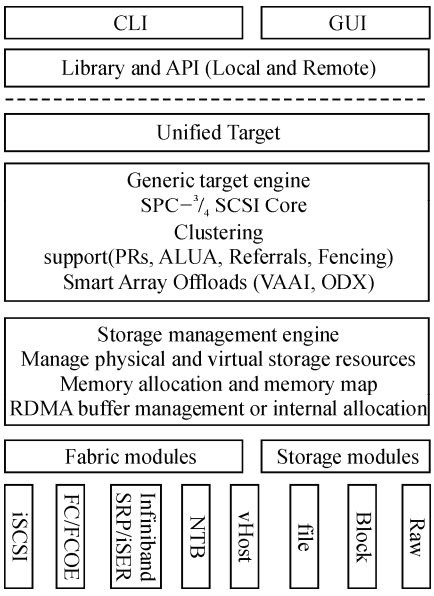


图 3　LIO 组成模块

图 3 为 LIO 的组成模块，它由虚线以上用户态的管理工具和虚线以下 linux kernel 里面的模块组成。用户态管理工具为 targetcli，通过该工具可以配置 target 端的后台存储设备和不同形式的 SCSI target。linux kernel 模块工作原理是 Storage Management Engine 屏蔽不同后台存储设备差异并提供统一的存储访问接口给 Generic Target Engine。Fabric Modules 处理 initiator 和 target 的通信，接收 initiator 发来的 SCSI 命令并发送给 Generic Target Engine，Generic Target Engine 执行 SCSI 命令并把结果返回给 initiator。

图 4 为 SCST 的组成模块，它由 SCST core、device handlers、target drivers 和 user space utilities 组成，实现方式是将 target 具体驱动和存储设备驱动封装成可注册的插件驱动。

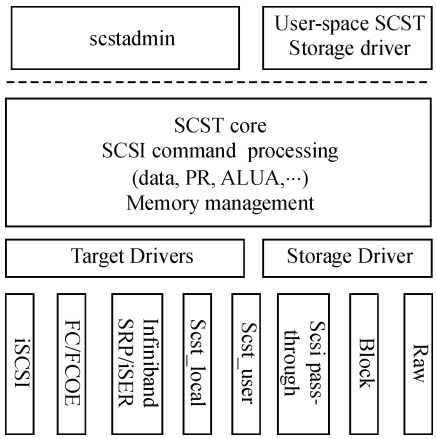


图 4　SCST 组成模块

## 2　解决方案

多协议网络存储系统的实现需要从硬件平台、软件支持和部署方式 3 个方面考虑。多协议网络存储系统属于 SAN 架构，其硬件平台和软件方案如下所述。

### 2.1　硬件平台选择

硬件方案在网络存储系统的组建中并不是关键，任何一台普通主机都能提供 SAN 存储服务，此处需要考虑的是数据在网络存储系统中的传输路径和该路径上影响数据传输速度的几个因素，网络存储系统中的 I/O 通路如图 5 所示。

从图 5 中可见：网络存储系统系统性能受到综合设计的影响，主要影响因素包括磁盘阵列端系统总线带宽、前端总线带宽、南桥上磁盘控制器带宽、主机端和磁盘阵列端的物理链路[7]，这几项带宽如表 1。

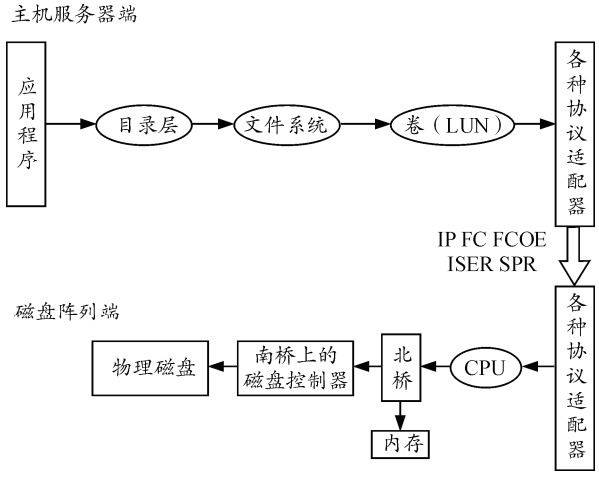


**图 5　网络存储系统中的 I/O 通路**

**表 1　网络存储系统硬件平台带宽**

| 影响因素 | CPU | DDR3 | 主板磁盘控制器 | PCIE2.0 | | FC | Infiniband |
|---|---|---|---|---|---|---|---|
| 频率/MHz | 3 100 | 1 333 | — | X4 | X8 | X4 | X8 |
| 带宽/(Gbit/s) | — | 10.664 | 0.3 | 2 | 4 | 0.5 | 7 |

从表 1 中可见：主板磁盘控制器和存储链路会成为整个存储系统的瓶颈，一般磁盘控制器带宽可以通过对多个物理磁盘同时访问来提高；所以目前提供存储系统性能一般需要提高存储链路的带宽，目前可以采用万兆网卡、光纤卡和 infiniband 卡来扩展。

总之，硬件平台的选择原则是综合考虑整个存储系统的性能，不能只追求某个部分的高性能，需要找出以上影响因素中的瓶颈，提高该处的性能。

多协议存储系统在磁盘阵列的前端预留 3 个 PCIE 插槽，通过在该处安装以太网适配器、光纤适配器或 infiniband 适配器来组建各种形式的存储网络。此种方式改进了存储系统网络链路造成的系统瓶颈，提高了系统性能和可扩展性。

## 2.2　软件方案

通过在主机上安装 Target 软件，就能将主机底层的磁盘空间以虚拟磁盘或实体磁盘等形式，透过各种协议挂载给其他主机使用，让通用型服务器提供基于各种协议的 SAN 存储服务。

target 端软件有 2 种配置方式，即 linux 内核自带的 LIO 模式和开源组织开发的 SCST，这 2 种模式都提供了跨多种协议的 SCSI Target 功能，不仅可提供 iSCSI Target 功能，还能支持 FC、FCoE、iSER 和 SRP 等多种存储网络协议，是多协议存储系统实现的关键。

initiator 端需要安装网卡、光纤卡，infiniband 卡驱动及 initiator 软件，windows 下采用自带的 initiator 软件，linux 下采用 open-iscsi。

图 6 所示为网络存储系统的软件处理流程。在 initiator 端通过 open-iscsi 将 target 连接，经过 bio->scsi cmd->iscsi/fc/ib_srp cmd->tcp/ip/fc/srp packet，最终通过各种协议发送到 target 端，target 端验证是否允许该 initiator 访问之后建立连接(login)之后，就能在客户端发现标准 SCSI 设备，然后就像使用本地设备一样对其进行读写访问。

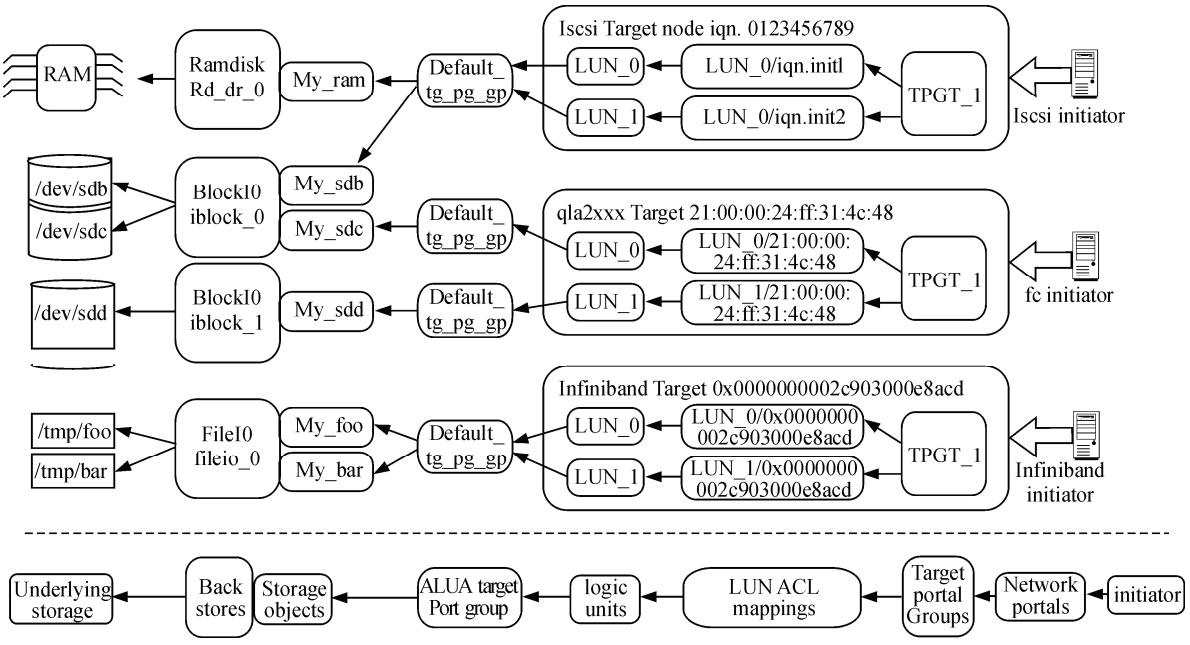


**图 6　多协议存储系统软件处理流程**

## 3　性能测试

### 3.1　软硬件环境

2 台普通 PC、2 个 QLE2460 光纤卡、2 个 Mellanox CX354A infiniband 卡。

测试软件：

1) 磁盘阵列端：ubuntu13.10（64 bit，内核版本 3.11.0）targetcli（LIO）、SCST2.2；

2) 主机服务器端：ubuntu13.10（64 bit，内核版本 3.11.0）：MLNX_OFED_LINUX-2.1-1.0.6-ubuntu13.10-x86_64.iso。

### 3.2　测试方法

将主机服务器端与磁盘阵列端分别采用以太网、光纤和 infiniband 链路互联，然后进行如下 2 种测试：1) 将磁盘阵列端的实际物理磁盘映射成 dev_disk 模式的后端存储设备，并映射到主机服务器端，监测从主机服务器端本地磁盘往远程虚拟磁盘拷贝大文件的速度；2) 在磁盘阵列端使用 dd 命令创建大小为 2 G 的/disk0 文件，并用作 vdisk_fileio 模式的后端存储设备，将其映射到主机服务器端。监测从主机服务器端内存往远程虚拟磁盘拷贝大文件的速度。

### 3.3　配置过程

磁盘阵列端采用 LIO 时，配置过程如下：

1) 重新配置内核选项并编译。

首先使能内核中自带的 target 核心模块 target_core.ko 和需要的后端驱动模块 iscsi.ko、tcm_fc.ko、loopback.ko，然后使能具体的 infiniband target 模块 srpt.ko 和用户管理及版本模块、Qlogic 光纤 target 模块 qla2xxx.ko 和底层 scsi 驱动，重新编译 linux 内核。

2) 安装管理软件 targetcli 并进行配置。

① 创建后端存储设备。

/backstores> iblock/ create name=my_disk dev=/dev/sdb

② 创建 target。

/backstores/iblock/my_disk> /ib_srpt create 0x0000000000000000002c903000e8acd

③ 创建 lun。

/ib_srpt/0x00...2c903000e8acd> luns/ create /backstores/iblock/my_disk

④ 定义权限。

/ib_srpt/0x00...2c903000e8acd> acls/ create 0x0000000000000000002c903000e8be9

磁盘阵列端采用 SCST 时，配置过程如下：

首先编译安装 scst 核心模块、target 模块和存储模块，然后安装 scst 管理工具 scstadm。scst 提供 scstadm、sysfs 和 procfs 3 种方式配置本地 target，具体过程类似于 LIO。

主机服务器端操作系统选用 ubuntu13.10，需要在内核选项中选中光纤卡及 infiniband 卡驱动，另外安装 open-iscsi。

### 3.4　测试结果

多协议存储系统传输带宽如表 2。

**表 2　多协议存储系统传输带宽**

| 连接方式 | 存储设备类型 | SCST | LIO |
|---|---|---|---|
| ISCSI | dev_disk/(Mbit/s) | 116 | 118 |
| | vdisk_fileio/(Mbit/s) | 86 | 120 |
| FC | dev_disk/(Mbit/s) | 127 | 218 |
| | vdisk_fileio/(Mbit/s) | 312 | 383 |
| Infiniband | dev_disk/(Mbit/s) | 223 | 246 |
| | vdisk_fileio/(Gbit/s) | 1.41 | 1.8 |

从表中可见：在对实际磁盘进行写操作时，除了受到千兆网络传输上限 125 Mbit/s 的限制之外，其余传输带宽都在 200 Mbit/s 左右，即南桥集成的 sata 控制器传输带宽的上限；在内存间进行数据写操作时，如果采用光纤连接，传输带宽接近 400 Mbit/s，除去一些软件开销，基本达到光纤通道带宽上限 500 Mbit/s，如果采用 infiniband 卡连接，传输带宽接近 1.8 Gbit/s，该值与 infiniband 卡的理论带宽 7 Gbit/s 还有很大差距，这是因为实验中采用 ramdisk 的方式将内存映射成虚拟磁盘，此种方式采用的 ramdisk 软件本身会模拟 scsi 控制器影响传输带宽；相同实验条件下，LIO 软件效率稍优于 SCST。

## 4　结束语

在对网络存储系统进行设计时，需要综合考虑整个 I/O 通路上的带宽并找出瓶颈，通过提高瓶颈处的性能来提高整个存储系统的性能。当前存储系统性能瓶颈一般出在磁盘阵列和主机端连接的网络链路上，可以通过光纤卡或 infiniband 卡代替网卡来提升该处性能。LIO 和 SCST 是磁盘阵列端可选的 2 种 scsi target 软件，支持 IP、FC、FCOE、SRP 和 iSER 多种协议，通过测试得出，LIO 性能优于 SCST，是多协议网络存储系统的首选。

（下转第 75 页）