第4章 存储协议及接口技术

4.1 项目导引——精益求精

“睿”解决了资料的安全存储问题后感到非常高兴，邀请了家族的亲朋好友一起来聚会庆祝这个成果。大家高兴之际，一个小孩有可能受到了“汉”家族优秀的基因的影响，非想搞明白这些老祖宗的思想是怎么在这些机器之间传递的。这下可把“睿”给难住了，他答应这个小孩，等下次聚会时一定告诉他这个问题的答案。“睿”为了弄懂这个问题，请教了远房的家族亲友——计算机专家“留洋”。这个“留洋”可是大有来头，在国外学习和研究计算机近 20 年，他是怎样为“睿”讲解的呢？“睿”能否学习明白呢(图 4-1)?

4.2 项目分析

“留洋”向“睿”分析了他所关心的问题，那些《本草纲目》、《黄帝内经》、《齐民要术》等都是存储在磁盘系统中的数据，我们要看到这些书，一定要把它从服务器的磁盘中读出来，然后用我们的电脑显示出来观看，而那个小孩关心的是这个过程是怎么做到的。这是数据的传输方式和处理方式的问题。磁盘的工作方式“睿”已经在之前研究过了，现在最想研究明白的就是磁盘是怎样把数据从服务器的磁盘传到“汉”的孩子用的电脑上的。其中最重要的就是用什么样的技术传输数据，以及这些技术是怎样工作的？这些技术包括存储接口标准和应用的协议。存储系统如图 4-2 所示。

图 4-1 精益求精

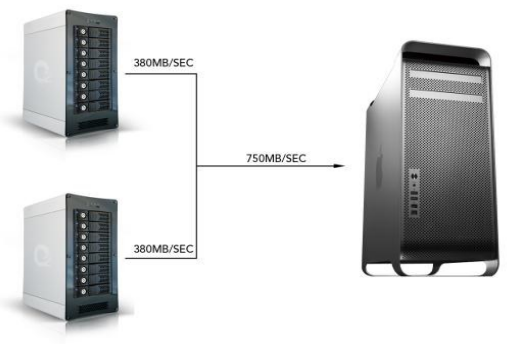

图 4-2 存储系统

4.3　技术准备

数据存储于磁盘中，其最终目的是通过传输与输出成为有用的信息，为人所用。而数据传输最重要的就是准确与高速。为了达到这两个要求，科技人员通过大量的实验和应用，为存储的数据 I/O 提出了 SCSI、FCP、iSCSI 等协议与接口标准。为数据进行准确、快速的 I/O 建设好了快车道。不同的协议应用于不同的存储结构类型，要以需求和应用的环境来取舍与配合使用。

4.3.1　SCSI 总线协议

SCSI 协议的主要功能是在主机和存储设备之间传送命令、状态和块数据。例如：有一个应用程序向操作系统发出对磁盘设备的写请求。在 SCSI 协议层，这个写请求被看成是特定数量的数据块以协议的形式传递到指定位置的命令。作为操作系统和存储设备之间的一个中介，SCSI 协议既不规定数据块如何组织，也不规定怎样把数据块放到磁盘上。在 SCSI 把数据块发送到目的地时，目标方可能是单个物理磁盘，也可能是把数据块在多个物理磁盘上分条存放的 RAID 控制器。SCSI 协议的责任，就是在确认写操作已经正确完成后向操作系统报告成功，而不管在磁盘上物理存储是如何配置以及写操作是如何执行的。SCSI 系统结构如图 4-3 所示。

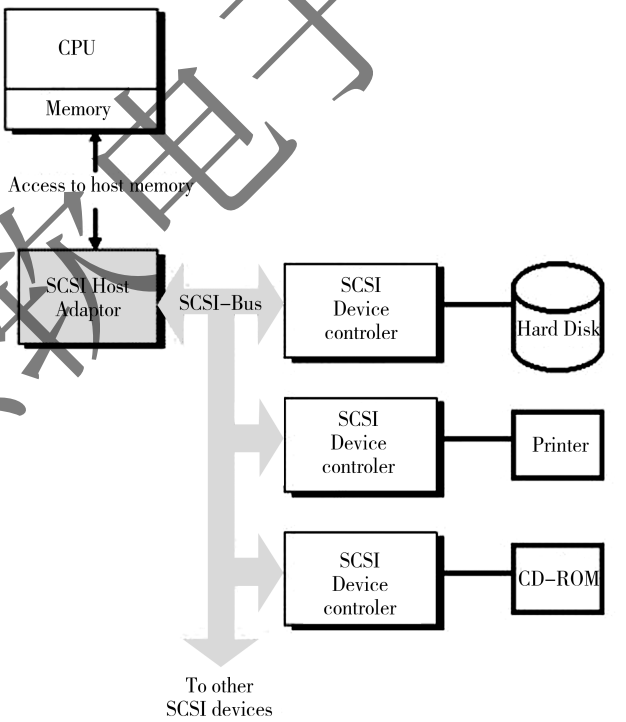

图 4-3　SCSI 系统结构图

由生成并发出请求命令的 SCSI 发起方、接受并处理此命令的 SCSI 目的方、数据交换的总线以及应用的 SCSI 设备可以构成一个 SCSI 域。而任何连接到此域中进行数据服务的域成员

设备都必须遵守 SCSI 协议。

1. SCSI 域的组成与设备

SCSI 域由 SCSI 的控制器、设备、电缆和终接器组成。SCSI 控制器也称为主机适配器，控制器既可以是插入可用插槽的卡，也可以内置在主板上，协调 SCSI 总线上的设备和计算机之间的数据 I/O。SCSI BIOS(Basic Input Output System)在控制器上，是一个固化在小型 ROM 或闪存芯片内的负责控制连接总线设备的程序软件。每个 SCSI 设备都必须具有唯一的标识符(ID)才能正常工作。例如，如果总线能够支持 16 个设备，通过硬件或软件设置指定的设备 ID 的范围为 0～15。SCSI 控制器本身必须使用其中一个 ID，通常是最高的那一个，而将其他 ID 留给总线上的其他 15 个设备使用。内部设备通过带状电缆连接到 SCSI 控制器。外部 SCSI 设备使用一条粗的圆形电缆，以菊花链形式连接到控制器(串行连接 SCSI 设备使用 SATA 电缆)。在菊花链中，每个设备都依次连接到下一个设备。因此，外部 SCSI 设备通常具有两个 SCSI 连接器——分别连接前后两个设备。

电缆本身通常由三层构成：

(1)内层：保护性最好的层，包含实际发送的数据。

(2)介质层：包含向设备发送控制命令的线路。

(3)外层：包含传输奇偶校验信息的线路，这些信息可确保数据的正确性。

不同 SCSI 标准使用不同的连接器，这些连接器通常不兼容，通常使用 50、68 或 80 针。SAS 使用较小的 SATA 兼容连接器。一旦总线上的全部设备安装完毕，而且分配了各自的 ID，则总线的每一端都必须闭合。

如果 SCSI 总线保持开放状态，沿总线发送的电信号会反射回来，从而干扰设备和 SCSI 控制器之间的通信。解决方法是终结总线，用电阻电路闭合每一端。如果总线同时支持内部和外部设备，则必须终结每个系列的最后一个设备。

SCSI 终结的类型主要可分为两类：被动(无源)和主动(有源)，如图 4-4 所示。

图 4-4　各种 SCSI 终结器

(1)被动(无源)终结通常用于在标准时钟速度下运行、且设备到控制器的距离小于 1 米的 SCSI 系统。

(2)主动(有源)终结用于 Fast SCSI 系统，设备到 SCSI 控制器的距离大于 1 米的系统。

SCSI 还使用三种不同类型的总线信令，这也会影响终结。电脉冲以信令的方式在线路上发送。

(1)单端(SE：Single-ended)：控制器生成信号，并通过单条数据线将信号传送至总线上的

所有设备。每个设备都会产生信号损失。因此，信号会很快开始衰减，由此 SE SCSI 的传输距离被限制为约 3 米以内。PC 中普遍采用 SE 信令。

(2)高压差动(HVD：High-Voltage Differential)：HVD 常用于服务器，它以串联方式发送信号，采用一条数据高压线和一条数据低压线。SCSI 总线上的每个设备都有信号收发器。控制器与设备通信时，总线沿途的设备接收信号并转发信号，直至信号到达目标设备为止。这样，控制器和设备之间的允许距离可显著增加，可达 25 米。

(3)低压差动(LVD：Low-Voltage Differential)：LVD 是 HVD 的同类技术，工作原理非常相似。两者之间的差异在于，LVD 的收发器更小，并且内置于每个设备的 SCSI 适配器中。这使得 LVD SCSI 设备的价格更合理，并且 LVD 使用更少的电量就可以通信。缺点在于最大距离仅为 HVD 的一半——12 米。

HVD 和 LVD 通常都使用被动终结器，即使设备和控制器之间的距离远大于 1 米也是如此。这是因为收发器可以确保信号足够强，能从总线的一端传输到另一端。

SCSI 协议定义设备间是一对一进行数据交换的，即采用分时总线共享方式工作。

2. SCSI 协议的分层结构

为了便于实现和理解 SCSI 协议，SCSI 采取了分层结构。包括 SCSI 应用层，SCSI 传输层和 SCSI 互连层。SCSI 中的各个具体协议一般都位于其中的某一层。

(1)应用层：SCSI 体系结构把发起方(主机)和目标方(如磁盘)的通信定义为客户/服务器交换。SCSI 客户位于主机中，代表上层应用程序、文件系统和操作系统 I/O 请求。SCSI 设备服务器位于目标设备中，对请求做出响应。客户/服务器请求和响应通过某种形式的底层协议进行传输。

(2)传输层：SCSI 设备之间通过一系列的命令实现数据的传送，大致分成三个阶段：命令的执行，数据的传送和命令的确认。

(3)互联层：完成 SCSI 设备对总线的连接以及发送方和目标方的选择等功能。

3. 一个 SCSI 目标的三元标识：总线/目标设备/逻辑单元号

为了对连接在总线上的设备寻址，SCSI 协议引入了 SCSI 设备 ID 和逻辑单元号 LUN. 在 SCSI 总线上的每个设备都必须有一个唯一的 ID，其中包括服务器中的主机总线适配器也要拥有设备 ID。这取决于 SCSI 标准的版本，每条总线最多可允许有 8 个或者 16 个设备 ID。

诸如 RAID 磁盘子系统和磁带库这样的存储设备可能包括若干个子设备，如虚拟磁盘，磁带驱动器和介质更换器等。因此 SCSI 引入了逻辑单元号，以便于对大的设备中的子设备进行寻址。另外一个服务器可能配置了多个 SCSI 控制器，从而就可能有多条 SCSI 总线。因此，操作系统用一个三元描述标识一个 SCSI 目标：总线/目标设备/逻辑单元号。

传统的 SCSI 适配卡连接单个总线，相应的只具有一个总线号。在引入存储网络之后，每个光纤通道 HBA(Host Bus Adapter)或 iSCSI(Internet SCSI)网卡也都连接一条总线，分配一个总线号，在它们之间依靠不同的总线号加以区分。

目标设备标识在一条总线菊花链上的单个设备上，逻辑单元号则表示一个目标设备中的一个子设备。通常，单个物理磁盘只具有一个逻辑单元号，而 RAID 磁盘阵列虽然也只有一个目标设备，但却有多个逻辑单元号。

在一条总线上各个设备具有不同的优先级。起初的 SCSI 协议只允许有 8 个目标设备 ID，

规定 ID7 具有最高权限。后来版本的 SCSI 协议允许有 16 个不同的目标设备 ID。出于兼容性的考虑，从 7 到 0 的目标设备依然具有高优先级，而从 15 到 8 的设备 ID 具有较低的优先级。

设备（服务器和存储设备）在可以通过 SCSI 总线发送数据之前必须预定总线（仲裁）。在总线的仲裁期间，具有最高优先权的目标设备总能获胜。在总线负载重的情况下，这可能导致具有较低优先级的设备总是不被允许发送数据，因此，SCSI 的仲裁过程是不平等的。

出于配置和管理的需要，操作系统使用总线号/目标设备 ID/逻辑单元号三元组来标识一个 SCSI 目标，然而用户和应用程序所看到的只是一个逻辑标识符，如 D 盘。因此在总线号/目标设备 ID/逻辑单元号和逻辑盘符之间存在着一个映射，提供在物理设备和上层文件系统之间不同表示形式的转换。

4. SCSI 协议的通信方式

SCSI 协议把发起方（主机）和目标方（例如磁盘）之间的通信定义为客户/服务器方式。应用客户位于主机中，代表上层应用程序、文件系统和操作系统的 I/O 请求。设备服务器位于目标设备中，它响应客户的请求。请求和响应通过某种形式的下层分布设施进行传输，该分布设施称作分布子系统，可以是并行电缆，也可以是光纤通道协议或 iSCSI。

一个发起方可能会有多个请求同时发给目标方。多个请求产生应用客户的多个实例，从而在设备服务器上产生多个事务。

发起方在其发往一个或多个目标的多个请求正在被相关的设备服务器处理的时候，需要能够执行上下文交换（Context Switching），即具有从一个任务快速切换到另一个任务的能力。例如，作为一个发起方的文件服务器可以向一个目标方发送一个写请求。当该文件服务器在等待这个目标方准备好缓冲区以接收数据的那段时间内，可以切换到另一个挂起的任务，例如处理已经到达的对先前的另一个请求的响应，从而提高运行效率，实现最大化吞吐量。如果 SCSI 任务只能依次串行地执行，那么等待每个写或读请求完成的时间就都被白白地浪费了。一般来说，上下文交换是由主机适配卡完成的，可以是并行 SCSI，也可以是光纤通道或 iSCSI。

由于 SCSI 体系结构模型是层次化的，因此它对主机 I/O 请求的处理可以独立于底层的分发子系统。一个应用客户主机可以处理涉及不同种类的目标设备的 I/O 操作，例如一个应用服务器可以有直接附接的 SCSI 目标方，也可以有通过千兆位速率接口连接的串行 SCSI 目标方。

在 SCSI 发起方和目标方之间读写数据是通过 SCSI 命令、分发请求、分发操作和响应来完成的。SCSI 命令和参数在 CDB（Command Descriptor Block，命令描述块）中指定。作为交互示例，在执行对磁盘的 SCSI 写过程时，在发起方（例如主机总线适配器）创建一个应用客户，该客户发送 SCSI 命令请求给目标方，令其准备缓冲区以接收数据。目标设备服务器在其缓冲区准备好之后，发送一个数据分发操作请求进行响应。接着，发送方就执行分发操作，开始发送数据块。依赖于底层的分发子系统，数据块可能按字节并行传输（例如并行 SCSI 总线），也可能以分段成帧的形式串行传输（例如光纤通道或 iSCSI）。

从应用程序或操作系统的角度看，写操作只是一个事务。但实际上，对应一个写操作，发送方和目标方可能要进行多次的分发请求和分发操作的交互，才能把命令请求的所有数据都发送给目标方。

在一次读操作中，SCSI 命令块遵循相反的数据分发请求和确认序列，然而由于是发起方发出读命令，所以命令就假定自己已经准备好了缓冲区以接收第一批数据块。在读写事务的每个

阶段所发送的数据块数量，由发起方和目标方根据对方的缓冲区容量协商决定。例如，高性能磁盘阵列一般都能提供较大的缓冲区，可以完成大规模的数据传送，从而提高了产品性能。

5. SCSI 的读操作和写操作过程

（1）SCSI 的读操作过程。

如果计算机要从存储设备上读取文件或数据，那么无论数据的大小如何，都至少要经历一个 SCSI 的读操作过程。当然，操作系统需首先将用户的读取操作通过 SCSI I/O 的应用程序编程接口（Application Programming Interface，API）转化为 SCSI 的读操作，并在操作完成后通过相应的 API 返回响应的值。

在 SCSI 域内，这个操作在传输层被简单地描述成 五个主要过程：

- 发起方通过 CDB 发送 SCSI 的读命令。
- 目标方接收到该命令，通过设备管理器在指定逻辑单元中执行该命令请求的操作。
- 目标方以字节为单位向发起方传送数据。
- 在数据传输完毕后，目标方向发起方发送命令完成的报告。
- 发起方接收到命令完成的响应。

当然，这些过程是建立在 SCSI 互连层的基础上的。在第一个过程之前，SCSI 总线由空闲阶段进入总线仲裁和选择阶段，完成发起方对总线使用权的获得以及对目标方的选择和寻址。

在第一个过程中，目标方发送 REQ 信号，请求信息传输，控制总线进入信息传送的命令阶段。目标方通过发送方传送的 CDB 获取“读”命令。在其后的第二个和第三个过程中，目标方从它控制的外围设备中读取数据并发送到发起方。如目标方准备数据需要较长的时间，则可能有多个总线释放、进入空闲和重选阶段的轮回。目标方在每次完成数据传送后，都控制总线进入状态阶段并返回一个状态信息。为进一步表示读命令的全部完成，在第四个过程中，总线进入信息传送的通信阶段，目标方发送“命令完成”信息，并可释放 SCSI 总线的 BSY 信号。在第五个过程中，发起方接收到目标方命令完成的响应，总线可恢复到空闲阶段。

（2）SCSI 的写操作过程。

SCSI 的写操作过程与读操作过程类似，但数据传送的方向不同，它把数据从发送方向目标方传送。在发送方系统中有对文件做写操作的用户请求时，它先通过文件系统查找该文件在存储设备（如磁盘）上的逻辑块地址（Logical BlockAddress，LBA），接着文件系统把该 LBA 连同其他一些参数，如数据的指针、数据的长度以及逻辑单元号等传递给 SCSI 的 API，并指示一个写操作。例如写 6000 字节到 LUN0 的逻辑块地址 0001234AB。SCSI 的 API 则具体发送一个写命令给 LUN0，并将数据以存储设备认可的方式分批或一次性地传递到 LUN0，直到数据全部传输完毕。之后，SCSI 的 API 返回，并指示任务完成。然后，文件系统通知应用程序任务完成。至此，一个文件的写操作完成。

当然，在数据写操作中，仍然需要具体运行 SCSI 的各个阶段，并需要发送 SCSI 信号以及 SCSI 命令，如写命令等。这些方面都与上面描述的读操作类似，此处不再赘述。

从上面的介绍可以看出，一个简单的数据读或写操作会涉及一系列的过程。实际上，在这些过程中，除了有应用程序（如字处理软件、数据库等）为用户提供的直接操作界面和操作系统给应用程序提供的通用的系统功能外，还有文件系统、SCSIAPI、SCSI 设备命令、SCSI 驱动程序、总线和存储设备等多种软硬件的参与。这就是 SCSI 协议的基本工作原理。

4.3.2 Fibre Channel 协议

FCP(Fibre Channel Protocol)是指光纤通道协议。网状信道标准(FCS)定义了一种用于连接工作站、大型机、巨型机、存储设备以及显示设备等的高速数据传输机制。FCS 满足了大量信息的快速传输需求,并减轻了支持当前多通道和网络环境的系统提供商的负担,这是因为 FCS 为网络、存储以及数据传输提供了一个单一标准。网状信道协议是一种在网状信道上的 SCSI 接口协议(图 4-5)。

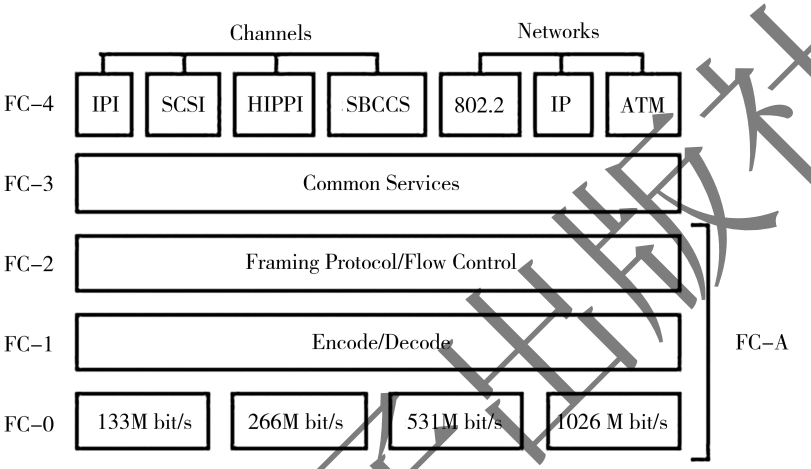

图 4-5 网状信道协议

网状信道(FC)主要特征如下:

- 性能:传输速率范围从 266Mb/s 到 4Gb/s;
- 支持光介质和电介质,工作速率范围从 133Mb/s 到 1062Mb/s;
- 传输距离为 10km;
- 小连接器;
- 高带宽利用率,与距离无关;
- 支持从小型机到巨型机的多种性价比级别;
- 可以传送多种接口命令集,包括 IP、SCSI、IPI、HIPPI-FP 以及音频/视频。

FC 各组成部分如下:

- FC-0:物理介质接口;
- FC-1:数据的编码和解码以及在物理介质上传输的物理链路控制信息;
- FC-2:传输构成帧、序列以及交换的协议信息单元;
- FC-3:提供诸如分段处理(striping)、节选组(hunt group)以及组播等高级特征所需要的通用服务;
- FC-4:运行在网状信道上的应用程序接口,如网状信道协议(FCP)中的小型计算机系统接口(SCSI)。

网状信道(FC)中的基本实体是网状信道网络,与一般分层网络不同的是,一个网状信道网络很大程度上由功能单元以及各单元间接口所指定,各部分组成如下:

- N_PORTs:网状信道流量终点;

- FC Devices：N_PORT 访问的网状信道设备；
- Fabric Port ：网状网络接口，连接 N_PORT ；
- 在 N_PORT 间传输数据帧的网络结构；
- 交换结构或混合结构下的一组辅助服务器，包括支持设备发现和网络地址解析服务的名称服务器。

主要的网状信道网络拓扑组成如下所示：

- Arbitrated Loop：N_PORTs 以菊花链(daisy-chain)形式连接在一起；
- Switched Fabric：由交换单元组成的网络；
- Mixed Fabric：由交换机和“fabric-attached”环路组成的网络。L_PORT 将 loop-attached N_PORT (NL_PORT)与环路连接起来，并且 NL_PORT 通过 FL_PORT 接入该结构。

4.3.3　iSCSI 协议

iSCSI(互联网小型计算机系统接口)是一种在 TCP/IP 上进行数据块传输的标准。它是由 Cisco 和 IBM 两家发起的，并且得到了各大存储厂商的大力支持。iSCSI 可以实现在 IP 网络上运行 SCSI 协议，使其能够在诸如高速千兆以太网上进行快速的数据存取备份操作。

iSCSI 标准在 2003 年 2 月 11 日由 IETF(互联网工程任务组)认证通过。iSCSI 继承了两大最传统技术：SCSI 和 TCP/IP 协议。这为 iSCSI 的发展奠定了坚实的基础。

基于 iSCSI 的存储系统只需要不多的投资便可实现 SAN 存储功能，甚至直接利用现有的 TCP/IP 网络。相对于以往的网络存储技术，它解决了开放性、容量、传输速度、兼容性、安全性等问题，其优越的性能使其备受关注与青睐。

1. iSCSI 的数据包结构

iSCSI 的数据包结构如图 4-6 所示。

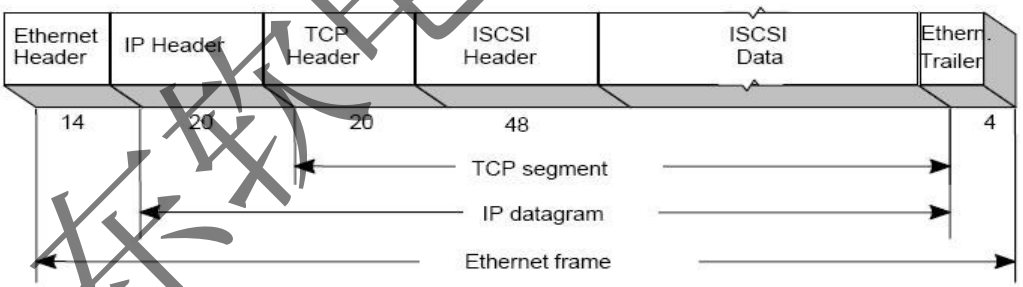

图 4-6　iSCSI 的数据包结构

2. 工作流程

- iSCSI 系统由 SCSI 适配器发送一个 SCSI 命令。
- 命令封装到 TCP/IP 包中并送入到以太网络。
- 接收方从 TCP/IP 包中抽取 SCSI 命令并执行相关操作。
- 把返回的 SCSI 命令和数据封装到 TCP/IP 包中，将它们发回到发送方。
- 系统提取出数据或命令，并把它们传回 SCSI 子系统。

3. 安全性描述

- iSCSI 协议本身提供了 QoS 及安全特性。

● 可以限制 initiator 仅向 target 列表中的目标发登录请求，再由 target 确认并返回响应，之后才允许通信。

● 通过 IPSec 将数据包加密之后传输，包括数据完整性、确定性及机密性检测等。

4. iSCSI 的优势

(1)广泛分布的以太网为 iSCSI 的部署提供了基础。

(2)千兆/万兆以太网的普及为 iSCSI 提供了更大的运行带宽。

(3)以太网知识的普及为基于 iSCSI 技术的存储技术提供了大量的管理人才。

(4)由于基于 TCP/IP 网络，完全解决数据远程复制(Data Replication)及灾难恢复(Disaster Recover)等传输距离上的难题。

(5)得益于以太网设备的价格优势和 TCP/IP 网络的开放性和便利的管理性，设备扩充和应用调整的成本付出小。

4.3.4 iSCSI 与光纤通道的比较

从传输层看，光纤通道的传输采用其 FC 协议，iSCSI 采用 TCP/IP 协议。

FC 协议与现有的以太网是完全异构的，两者不能相互接驳。因此光纤通道是具有封闭性的，而且不仅与现有的企业内部网络(以太网)接入，也与其他不同厂商的光纤通道网络接入(由于厂家对 FC 标准的理解的异样，FC 设备的兼容性是一个巨大的难题)。因此，对于以后存储网络的扩展由于兼容性的问题而成为了难题。而且，FC 协议由于其协议特性，网络建完后，加入新的存储子网时，必须要重新配置整个网络，这也是 FC 网络扩展的障碍。

iSCSI 基于 TCP/IP 协议，它本身就运行于以太网之上，因此可以和现有的企业内部以太网无缝结合。TCP/IP 网络设备之间的兼容性已经无需讨论，迅猛发展的 Internent 网上运行着全球无数家网络设备厂商提供的网络设备，这是一个最好的佐证。

从网络管理的角度看，运行 FC 协议的光网络，其技术难度相当大。其管理采用了专有的软件，因此需要专门的管理人员，且其培训费用高昂。TCP/IP 网络的知识通过这些年的普及，已有大量的网络管理人才，并且，由于支持 TCP/IP 的设备对协议的支持一致性好，即使是不同厂家的设备，其网络管理方法也是基本一致的。

FC 运行于光网络之上，其速度是非常快的，现在已经达到了 2G 的带宽，这也是它的主要优势所在。下一代的 FC 标准正在制定当中，其速度可以达到 4G。

今天的千兆以太网已经在普及当中，这也是基于 TCP/IP 的 iSCSI 协议进入实用的保证。得益于优秀的设计，以太网从诞生到现在，遍及了所有有网络的地方，到现在依然表现出非凡的生命力，在全球无数网络厂商的共同努力下，以太网的速度稳步提升，千兆网络已经实际应用，万兆网络呼之欲出，以太网的主要部件交换机、路由器均已有万兆级别的产品。随着产品的不断丰富，以及设备厂商间的剧烈竞争，其建设成本在不断下降，万兆网络的普及已日益临近。当 iSCSI 以 10Gb 的高速传输数据时，基于 iSCSI 协议的存储技术将无可争议的成为网络存储的王者。

4.4　项目实施

经过学习，“睿”学懂了磁盘输入输出所应用的接口技术，开始认真的准备着下一次与那个小孩的交流。能不能给小朋友讲明白就看“睿”的本事了。

4.5　技术拓展

4.5.1　SCSI 总线信号

SCSI 在物理信号的基础上定义了一组总线信号。这些信号可划分为数据信号和控制信号两类。它们都是二进制信号，并且只有“真”和“伪”两个稳定状态。其中有指示总线已经被占用的“BSY”信号，有清除并重新设置 SCSI 总线的“RST”信号等。

下面对这些信号的名称和功能逐一进行介绍。

(1) BSY (Busy，忙)信号。

该信号是“或态”信号，表示已经有设备占用总线。

(2)SEL (Select，选择)信号。

该信号是“或态”信号，由发起方用以选择目标方，或者由目标方用以重新选择发起方。

(3) C/D (Control/Data，控制/数据)信号。

该信号由目标方驱动，表示在数据总线上传送的是数据信号还是控制信号。该信号处于“真”状态时表示控制信号。

(4) I/O(Input/Output，输入输出)信号。

该信号由目标方驱动，控制数据在数据总线上的移动方向。当 I/O 信号为“真”时表示是对发起方的输入，数据由目标方向发起方传送；若 I/O 信号为“伪”，则表示数据由发起方向目标方传送。该信号也被用来区分选择和重选阶段。

(5)MSG (Message，通信)信号。

该信号由目标方驱动，表示总线处于信息传送的通信阶段。

(6) REQ (Request，请求)信号。

该信号由目标方驱动，表示有信息传输请求，请求一个 REQ/ACK 数据传送握手过程。

(7) ACK (Acknowledge，应答)信号。

该信号由发起方驱动，表示对 REQ 信号的应答。

(8) ATN (Attention，提醒)信号。

该信号由发起方驱动，指示一个提醒信息，表明发起方有一个消息要给目标方发送。

(9) RST (Reset，重置)信号。

该信号是“或态”信号,表示一个硬件重置状态,指示总线进入重新设置阶段,清除所有使用总线的 SCSI 设备。

(10) DB (DataBus,数据总线)信号。

DB 信号有两种,分别是用于 8 位数据总线的 DB (7～0,P)和用于 16 位数据总线的 DB (15～0,P)。这些信号都用于传送信息的值,它们包括数据比特信号,加上奇偶检验比特信号。

由于在 SCSI 总线上挂有多个设备,一些控制信号可能同时被多个 SCSI 设备驱动。这些信号被称作“或态”信号。对于“或态”信号,SCSI 设备不会主动将其驱动成“伪”,而是依赖总线终结器,在总线上的所有设备都没有驱动该信号时将其设置成“伪。”只要有 1 个或多个设备驱动该信号,该信号就是“真”。与“或态”信号相对照的是“非或态”信号。对于“非或态”信号,SCSI 设备可以将其驱动成“伪”。

4.5.2 总线的使用状态

根据对总线不同的使用,可以把 SCSI 总线状态划分成八个不同的阶段:空闲阶段,仲裁阶段,选择阶段,重选阶段,命令阶段,数据阶段,状态阶段和通信阶段。其中命令阶段,数据阶段,状态阶段和通信阶段都设计信息在总线的传送,所以又称为信息传送阶段。除了空闲阶段外,其他阶段的总线都被 SCSI 设备占用。

(1)总线空闲阶段。

总线空闲表明没有一个设备在使用 SCSI 总线,也表示在此状态下,SCSI 设备如果需要,可以使用总线。SCSI 设备需要在总线上的 SEL 信号和 BSY 信号都是“伪”之后,才可以检测总线是否处于空闲状态。

作为例子,SCSI 总线可能在下列情况下进入空闲状态:

- RST 信号被设置;
- 不成功的总线选择或重选;
- 目标设备解除连接;
- 目标设备命令完成。

一旦一个 SCSI 设备确定总线处于空闲阶段,它就可以申请总线仲裁,从而进入仲裁阶段。

(2)总线仲裁阶段。

在 SCSI 总线上的设备必须先获得总线连接权,然后才可以进行其他的操作。在默认条件下,看似挂在总线上的设备在逻辑上是与总线断开的,没有参与总线上的活动。SCSI 设备只有在需要进行数据传输和设备状态报告时才会申请总线连接权。SCSI 设备一旦得到了总线连接权,就将在发起方和目标方之间形成一个物理连接的通道,然后就可以进行数据传输了。

一般情况下,总线的获取与对目标方的选择都由发起方完成。为了更加高效地使用总线,在某些情况下,例如在有较长时间的 CPU 处理等待或设备存取等待时,需要释放总线以供其他设备使用。在等待的相关任务完成后,再重新进行总线仲裁和连接权获取操作,以继续进行暂停的工作。因此,有时目标方也可以执行总线操作和连接权获取操作,准确地讲,是再获取

操作。

SCSI 总线上的设备的优先级是由它的地址即 SCSI ID 决定的。在窄 SCSI 中的 ID 范围是 0～7，对应的优先级是从 1 到 8。在宽 SCSI 中的 ID 范围是 0～15，其中对应 ID0～7 的优先级是从 9 到 16 递增，而对应 ID8～15 的优先级是从 1 到 8 递增。在这里，我们用较大的数值表示较高的优先级，因此，ID7 具有最高优先级。在窄 SCSI 中，ID 0 具有最低优先级；在宽 SCSI 中，ID8 具有最低优先级。

SCSI 总线上的 ID 数目是与 SCSI 数据总线宽度一致的，因此，窄 SCSI 有 8 个 ID，宽 SCSI 有 16 个 ID。在窄 SCSI 中的 8 根数据线的编号是从 0 到 7，在宽 SCSI 中的 16 根数据线的编号是从 0 到 15。有趣的是，具有某个编号的数据线上的信号，还被用来表示具有对应号码 ID 的 SCSI 设备是否在执行选择或相关操作。例如当数据总线中的数据线 DB (2)在某个特定的阶段被驱动成真时，就可以表示其 ID 为 2 的设备已经在总线上执行了选择或相关操作。

在 SCSI 域中，主机是存储设备的主要使用者，且对存储的响应要求较高，因此通常主机的优先级最高，其分配的 ID 值也最大，在窄 SCSI 中是 ID 7，在宽 SCSI 中是 ID15。

总线仲裁就是在可能同时有多个设备请求的情况下，最终只给予其中的一个 SCSI 设备总线控制权的过程。SCSI 设备在检测到“总线空闲”并等待一个时延后即可以把总线置成 BSY，并把与它的 SCSI ID 对应的数据线信号置为“真”，开始总线仲裁申请。

在等待一个时延后，该 SCSI 设备需检测在数据总线上是否有更高优先级的 SCSI ID 也为“真”。如果总线上确有更高优先级的设备在进行总线申请，则该 SCSI 设备不再置 BSY 和对应的数据线为“真”，而会放弃总线仲裁申请，直到下一次“总线空闲”；否则，该设备就获得了总线控制权，并由该设备把 SEL 信号置为“真”。同时，总线上的其他 SCSI 设备则检测到 SEL 信号为“真”后，不再置 BSY 信号和对应的数据线为“真”，放弃总线仲裁申请。为了保证确实已经获得了总线控制权，该设备在置 SEL 信号为“真”后、传送其他信号前，需要有一定的时延。

在总线仲裁阶段结束时，总线上有 BSY、SEL 和与获得总线的 SCSI 设备的 ID，其对应的数据线的信号为“真”。

(3)选择阶段。

在选择阶段，得到总线使用权的 SCSI 设备在总线上选择目标设备，以便随后可以向该目标设备发送诸如读和写这样的命令。这个阶段主要是完成对具有特定 SCSI ID 的设备的选择，其相关协议的定义主要是在 SCSI 体系结构的互连层。需要注意的是，逻辑单元号 LUN 的寻址是逻辑单元通过 SCSI 传输层协议完成的，不在互连层。与 LUN 编址相关的协议在传输协议层描述。

赢得仲裁的 SCSI 设备在把 BSY 和 SEL 信号置成“真”，经过一小段时延后，即可进入选择阶段。作为发起方，赢得仲裁的 SCSI 设备不可以把 I/O 信号置成“真”。在此阶段，发起方需要把与自己的 SCSI ID 对应的数据线的信号和对应所要选择的目标设备的 SCSI ID 的数据线的信号置成“真”，经过一小段时延，再把 BSY 信号置成“伪”，然后等待目标方的响应。

例如，SCSI ID 为 6 的主机把对应自己的 ID 的数据线 DB(6)和对应目标设备的 ID(＝6)的

数据线 DB(0)置成“真”后,数据总线上信号值的状态将如下所示。

DS (0) DS (1) DS (2) DS (3) DS (4) DS (5) DS (6) DS (7)

1 0 0 0 0 0 1 0

此时,只有两个数据线的信号值是“真”。如果有多于两个的数据线为“真”,则目标方认为有误。目标方在 SEL 和对应它的 ID 的数据线的信号为“真”并且 BSY 和 I/O 信号为“伪”的情况下,就可以确定它自己已经被选为目标设备。此时,目标方设备应该重新把 BSY 信号置成“真”。发起方在检测到 BSY 为“真”的信号后,就把 SEL 信号置成“伪”。特别需要注意的是,在该阶段结束时,BSY 信号是由目标方置位的。

(4)重选阶段。

在 SCSI 目标设备忙于处理其内部事务(通常是执行对存储数据的读或写操作)期间,它可以在等待操作(比如把存储在设备中的数据读入缓冲区或把暂存在缓冲区的数据写入缓冲区)完成时释放总线供其他设备使用,并在操作完成后重新申请对总线的使用权。因此,重选阶段也发生在“总线仲裁阶段”之后。但与选择阶段不同,重选阶段由目标方启动,重新建立由发送方启动成功但被目标方挂断的连接。

在目标设备释放了总线之后,BSY 和 SEL 信号处于被置成“真”的状态。此时目标设备通过把 I/O 信号置成“真”使自己成为赢得对总线使用权的一方。在重选阶段,目标方也需要把与自己的 SCSI ID 对应的数据线的信号和对应发送方设备的 SCSI ID 的数据线的信号置成“真”,经过一段短的时延,再把 BSY 信号置成“伪”,然后等待发起方的响应。

发起方在 SEL、I/O 和对应它的 ID 的数据线的信号为“真”并且 BSY 为“伪”的情况下,就可以确定它自己已经被重选。被重选的发起方可以通过查看数据总线来验证重迭的目标方的 SCSI ID。然后,发起方设备重新把 BSY 信号置成“真”。目标方在检测到 BSY 为“真”的信号后,它也执行把 BSY 驱动成“真”的操作,并把 SEL 信号置成“伪”。

被重选的发起方在检测到 SEL 信号为“伪”后,就把 BSY 置成“伪”,而目标设备则继续把 BSY 设置成“真”,直到它放弃对总线的使用权为止。这样,在该阶段结束时,信号的状态与选择阶段一样,也是由目标方设置的 BSY 信号。

(5)信号传送阶段(命令阶段,数据阶段,状态阶段和通信阶段)。

命令阶段、数据阶段、状态阶段和通信阶段被组合在一起作为信息传送阶段,因为它们都被用来通过数据总线传送数据或控制信息。SCSI 使用 C/D、I/O 和 MSG 信号区分不同的信息传送阶段以及对应的信息传输方向。目标方驱动这三个信号,控制从一个阶段到另一个阶段的转变。发起方可以通过把 ATN 信号置成“真”请求一个“通信出”阶段,而目标方可以通过释放 MSG、C/D、I/O 和 BSY 信号引入总线空闲阶段。信息传送阶段使用一个或多个 REQ/ACK 握手过程控制信息传送。每个 REQ/ACK 握手过程允许传送一个或多个字节的信息。因为信息传送阶段一定是在选择阶段或重选阶段之后,所以不改变 BSY 和 SEL 信号。事实上,在该阶段,BSY 信号持续为“真”,SEL 信号持续为“伪”。

表 4-1 示出了 MSG、C/D 和 I/O 信号值与阶段名及信息传输方向之间的关系。其中的“出”和“入”是相对子发送方设备而言的,且数据传输方向由 I/O 信号确定。

命令阶段允许目标方请求发起方传送命令信息。在命令阶段的 REQ/ACK 握手过程中,目标方把 C/D 信号置成“真”,把 I/O 信号和 MSG 信号置成“伪”。

数据阶段包括“数据入”阶段和“数据出”阶段。

●“数据入”阶段允许目标方请求把数据从目标方传送给发起方。在“数据入”阶段的 REQ/ACK 握手过程中，目标方把 I/O 信号置成“真”，把 C/D 信号和 MSG 信号置成“伪”。

●“数据出”阶段允许目标方请求把数据从发起方传送到目标方。在“数据出”阶段的 REQ/ACK 握手过程中，目标方把 C/D 信号、I/O 信号和 MSG 信号都置成“真”。

表 4-1　　MSG、C/D 和 I/O 信号值与阶段名及信息传输方向之间的关系

MSG	C/D	I/O	阶段	具体阶段	传输方向
1	0	0	* （未用）		
1	0	1	* （未用）		
1	1	0	通信	通信出	从发送方到目标方
1	1	1	通信	通信入	从目标方到发送方
0	0	0	数据	数据出	从发送方到目标方
0	0	1	数据	数据入	从目标方到发送方
0	1	0	命令		从发送方到目标方
0	1	1	状态		从目标方到发送方

* 注：0＝伪，1＝真，* ＝保留未来定义

状态阶段允许目标方请求把状态信息从目标方传送给发起方。在状态阶段的 REQ/ACK 握手过程中，目标方把 C/D 信号和 I/O 信号置成“真”，把 MSG 信号置成“伪”。

通信阶段可以是“通信入”阶段或“通信出”阶段。无论是在“通信入”阶段，还是在“通信出”阶段，都可以传送多条消息。传送的第一个字节可以是单字节消息，也可以是多字节消息的首字节。在一个通信阶段可以传送多个多字节消息。

“通信入”阶段允许目标方请求把消息从目标方发送给发起方。在“通信入”阶段的 REQ/ACK握手过程中，目标方把 C/D 信号、I/O 信号和 MSG 信号都置成“真”。

“通信出”阶段允许目标方请求把消息从发起方传送到目标方。目标方在响应发起方建立的提醒条件时调用“通信出”阶段。在“通信出”阶段的 REQ/ACK 握手过程中，目标方把 C/D 信号和 MSG 信号置成“真”，把 I/O 信号置成“伪”。

4.5.3　同步传输与异步传输

与传统网络的数据包传送方式不同，SCSI 基于 REQ/ACK 信号控制数据传输的过程。根据 REQ 和 ACK 信号控制与数据总线置位时间的差别，信息传输又可分为异步传输和同步传输两个列别。而且，无论传输的方向如何，信息的传输都是由 REQ 信号开始，并且 REQ 信号都是由目标方控制和发送的。

(1)异步信息传输。

异步传输方式可用于数据阶段的数据传输，也可用于命令、状态和通信阶段的信息传输。首先，信息传输的方向是由 I/O 信号决定的。如果 I/O 信号为“真”，那么信息是由目标方向发起方传输。在此情况下，为了传送信息，目标方先把数据线 DB(7/15～0，P)信号置成对应想要传送的二进制数位序列的值，然后把 REQ 信号置成“真”。发起方在检测到 REQ 为“真”时，读

取数据总线的值，然后把 ACK 信号置成“真”。当目标方检测到 ACK 为“真”时，就可以改变或取消放置在数据总线上的值，并把 REQ 置成“伪”。发起方在检测到 REQ 置成“伪”时把 ACK 也置成“伪”。当目标方检测到 ACK 为“伪”时，总线上就完成了一次数据传输，并可进行下一次数据传输。

在异步传输方式中，每个 REQ/ACK 握手过程传送一个(对于窄 SCSI)或两个字节(对于宽 SCSI)的信息。特别需要注意的是，在此方式中，目标方在置 REQ 信号后，必须持续地把数据线 DB (7/15～0，P)置成对应所要传送的二进制数位序列的值，直到它检测到 ACK 为“真”为止。

如果 I/O 信号为“伪”，那么信息是由发起方向目标方传输。在此情况下，目标方通过把 REQ 置成“真”来请求信息。发起方驱动 DB (7/15～0，P)到它需要发送的二进制数位序列的值，然后把 ACK 置成“真”。此后，继续把 DB (7/15～0，P)信号置成这个二进制数位序列的值，直到 REQ 变成“伪”为止。目标方则是在检测到 ACK 变成“真”时，读 DB (7/15～0，P)的值，然后把 REQ 置成“伪”。发起方在检测到 REQ 变成“伪”时，它可以改变或取消放置在数据总线上的值，并把 ACK 置成“伪”。

此后，目标方可以通过把 REQ 置成“真”，继续请求信息。

(2)同步数据传输。

同步数据传输只在数据阶段使用，并且是在目标方和发起方之间建立同步数据传输协定之后使用。

与异步传输中的规则相同，当 I/O 信号为“真”时，数据是由目标方向发起方传输。目标方先把数据放置到数据总线上，即置 DB (7/15～0，P)对应的线路，然后把 REQ 置成“真”。在同步数据传输中，目标方在把 REQ 置成“真”后，需要把放置在 DB (7/15～0，P)上的二进制数位序列的值保持一个指定长度的时间，但不必维持到对 ACK 信号变“真”的接收。这是与异步传输不同的一个地方。在指定长度的时间期满后，目标方就可以把 REQ 置成“伪”，并且可以改变或取消放置在数据总线上的值，然后准备发送下一个数据。发起方在检测到 REQ 变“真”之后一个指定长度的时间内读 DB (7/15～0，P)上的值，然后把 ACK 置成“真”作为对目标方的响应。

与异步传输一样，在同步数据传输中，发起方也在接收到一个 REQ 并读取了数据总线上的值之后就发送一个 ACK 信号。但与异步传输不同的是，目标方在接收到对一个数据的 ACK 之前可以发送多个 REQ 信号。SCSI 为同步数据传输的 REQ/ACK 握手过程定义了一个称作 REQ/ACK 饱和值的参数，它表示在接收到 ACK 信号前可以发送的最大 REQ 信号数。如果发送的 REQ 数目多于接收到的 ACK 数目，并达到了定义的 REQ/ACK 饱和值，那么目标方暂停发送 REQ 信号和数据，直到接收到下一个 ACK 为止。这在原理上与传统网络中的流控制类似。

当 I/O 信号为“伪”时，数据是由发起方向目标方传输。发起方每次接收到一个 REQ 信号就发送一次数据。目标方先把 REQ 置成“真”。发起方检测到 REQ 变“真”后把要发送的数据放置到数据总线上，即置 DB (7/15～0，P)对应的线路，然后把 ACK 置成“真”。接着发送方需要在一个指定长度的时间内保持在总线上放置的数据不变，并继续把 ACK 置成“真”。在指定的时间期满后，发起方可以把 ACK 置成“伪”，并且可以改变或取消放置在数据总线上的值。

目标方在检测到 ACK 信号变“真”后，在指定的 ACK 保持为“真”的时间内读取数据总线上的数据，并把 REQ 置成“伪”。

此后，目标方可以通过把 REQ 再置成“真”继续请求信息。

4.5.4　SCSI 命令描述

在互连层完成 SCSI 设备对总线的连接，以及发送方和目标方的选择的基础上，传输层协议执行实际的数据传输。传输层提供了两类服务，一是命令的执行和确认；二是数据的传送。命令的执行是在总线进入命令阶段后，发起方通过命令描述块(command description block，CDB)向目标方发送具体的命令。命令的确认是在总线进入通信 (Message)阶段后，发起方接收由目标方发送的命令执行确认信息。数据的传送则是在数据阶段 (数据出或数据入)进行的。传输协议的运行过程包括发送命令、传输数据和对命令执行的确认。SCSI 基础命令规范 SPC (SCSI Primary Commands，SCSI 基础命令)定义了 CDB 的标准。

除了基本命令外，SPC 还定义了所有类型的 SCSI 目标方设备都可以使用的管理参数，如诊断参数和日志参数等。

发起方对存储设备的实际操作是通过向目标方发送一个命令描述块来完成的。在一些情况下，在一个命令描述块之后可能还有一些参数要传给目标方，按照具体的协定，这些更多的参数是在命令描述块后的“数据出”阶段发送的。命令描述块有定长和不定长两种格式，而定长格式的命令描述块又有 6、10、12 或 16 字节不同的长度规定。

命令描述块由编号从 0～5 的 6 个字节组成。下面介绍其中各个段的内容。

(1)操作码。

操作码是所有命令描述块都有的，它总是被放在命令描述块的开头一个字节。正如其名字所言，操作码定义 CDB 的具体操作。8 比特在理论上共有 256 个可能的操作码。实际上其中有一些是保留码，目前尚未定义。操作码的 8 个二进制位又分为两部分：5～7 位是组代码，指示该命令具体属于哪个命令组，它决定 CDB 的长度，如“000”为组“0”，表示 6 个字节的 CDB 命令组，0～4 位则是具体的命令代码。

(2)混杂 CDB 信息。

该参数表示与具体的 CDB 相关的信息，其中一个例子是表示逻辑设备号，寻址在 SCSI 目标设备中的一个逻辑单元。对应一个 SCSI ID 的设备可以有多个逻辑单元，所以逻辑单元扩展了 SCSI 总线可访问的设备数目，使得目标方设备上可以有多个可被访问的设备而只占用一个有效的 SCSI ID。对一个逻辑单元的实际访问是通过该逻辑单元的一个特定的编号，即逻辑单元号实现的。

(3)逻辑块地址。

该地址是逻辑单元 (比如磁盘)中的起始操作块的位置。在 6 字节的 CDB 中，有 21 位的逻辑块地址。SCSI 把逻辑单元、卷或分区抽象成块的数组，每一块都有一个逻辑地址，编号从 0 开始。对 SCSI 存储设备的每一次读/写操作都是针对一组连续的逻辑块进行的，因而需要指出起始块的逻辑地址。

(4)传送长度。

该长度表示命令所请求的传送量，通常是块数。在有些类别的 CDB 中也可能是字节数。0

表示不需要传送数据。

(5)参数表长度。

有些命令还需要更多的参数,这些参数由客户提供,定义在“数据出”缓冲区中。参数表长度就表示需要传送到存储设备的这类参数的长度,0 表示不需要传递参数。

(6)分配长度。

分配长度表示应用客户为“数据入”缓冲区分配的最大长度,根据具体的 CDB 类别,可能是字节数,也可能是块数。应用客户通常使用该“数据入”缓冲区接收特殊信息,如日志数据、诊断数据等。如果传送的信息量超过了分配长度表示的最大值,则相关设备不应再传,并使用状态阶段返回特定的状态信息。

(7)控制码。

它是所有 CDB 格式的最后一个字节。在其中有一些特殊的域,如已经定义的一个 NACA 位。在一些情况下,一个命令的执行会以“检查条件(Check Condition)”状态中止,它表明在命令执行过程中出现了错误或异常。有些命令执行的错误或异常不会影响其他命令的执行,也不需要作善后的恢复处理,而另一些命令执行的错误或异常则可能导致命令组中的其他命令被异常中止,需要专门的命令对其做善后处理,并要求存储设备在完成善后处理工作之前不再处理该用户的其他命令。为了区分这两种不同的情况,也为了让应用客户能够事先声明哪些命令执行的错误或异常需要善后处理,SCSI 允许应用客户在 CDB 的控制码中设置 NACA 位,请求存储设备在命令执行以“检查条件”状态中止时建立“自动跟随”条件(Condition),从而允许应用客户在随后的善后处理命令中把新(New)任务的属性设置成自动跟随(Auto Contingent Allegiance,ACA)。

4.6 本章小结

数据流是按规则和约定有序流动的,并非随机处理。SCSI 作为一种分时共享总线协议常用于 DAS 系统中,常见于服务器直连存储的结构。FCP 协议则是通过光传播的方式建立长距离、高速的存储交换网,常见于 SAN。而 iSCSI 则是通过依附于 TCP/IP 的网络协议,将数据传输进行了一次高性价比的移植。它的实施成本和使用效果优异,并且伴随着网络技术的升级换代,网络提速将赋予 iSCSI 更强大的性能和生命力。

4.7 拓展练习

拓展项目 1:SCSI 的技术特点和应用特点。

拓展项目 2:FCP 协议的技术特点和应用特点。

拓展项目 3:iSCSI 协议的技术特点和应用特点。