Broadview
www.broadview.com.cn

存储技术原理分析

基于Linux 2.6内核源代码

敖青云 著

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

存储技术原理分析

——基于Linux 2.6内核源代码

敖青云 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书通过对 Linux 2.6 内核源代码的分析，详细讨论存储技术的内在实现原理。分为三条主线：解释 PCI 设备、SCSI 设备及块设备的发现过程；跟踪存储 I/O 路径，即用户对文件的读/写请求怎么通过中间各个层次，最终到达磁盘介质；此外，还简要介绍主机适配器、块设备驱动及文件系统等编程框架。

书中将设计一些主要的场景，跟踪实现的各个层次，对其中的主要函数进行代码级的讲解。在分析每个模块时，会给出整体框架与主要数据结构之间的关系，并列出各个域的详细含义。

采用这种方式，希望读者能对存储相关概念（如 RAID、快照等）的内在实现有具体的了解，也试图帮助读者理解 Linux 内核设计和开发的一些思想，为进一步分析其他模块（如进程管理、内存管理等）起借鉴作用。

本书适合作为高校计算机相关专业本科生和研究生学习操作系统的辅助和实践教材，也适合作为 Linux 爱好者学习内核的参考书籍。同时，它也是存储从业工程师深入理解存储架构，以及软件开发工程师掌握软件架构的有效工具。

图书在版编目（CIP）数据

存储技术原理分析：基于 Linux 2.6 内核源代码 / 敖青云著. —北京：电子工业出版社，2011.9
ISBN 978-7-121-14432-5

Ⅰ.①存… Ⅱ.①敖… Ⅲ.①数据存贮 ②Linux 操作系统 Ⅳ.①TP333 ②TP316.89

中国版本图书馆 CIP 数据核字(2011)第 171290 号

责任编辑：许　艳
特约编辑：顾慧芳
印　　刷：北京天宇星印刷厂
装　　订：三河市皇庄路通装订厂
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：860×1092　1/16　印张：49.75　字数：1450 千字
印　　次：2011 年 9 月第 1 次印刷
印　　数：3000 册　定价：118.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前　言

笔者寄语

我终于长长地嘘了一口气！诺大的空间显得格外安静，稀疏坐落着的人们还在埋头忙碌着。此刻的我，压抑住内心喜悦，轻轻地踱到窗前，透过厚厚的玻璃，看着雪花乱舞，或覆盖车顶，或湮没河面，或消失在那一片新搭起的葡萄架下的泥土里，自己的思绪竟随着跳动起来。

无从知晓，写这本书的念头一直充斥着我的脑海，多年来，片刻都没有改变。为了你，我不惜被放逐天际，只要求拥有支配自己时间的自由；为了你，虽不能够遁入山林，我也坚持做着这大都会的隐者——浮华非我所求，喧嚣与我无缘；为了你，是的，我什么都愿意。

人生有如登山。源于上帝赐予的机遇，以及个人的努力，我有幸能登上这么一个小小的山峰，途中的辛酸只能自己去体味，或者最多和身边的朋友倾诉，但一路过来的心得，以及登临绝顶后所看到的风景却实在不敢独享。既然喜欢文字作画，为何不将它们记录下来，权当作同样有志于攀登的后来者的攻略，或者是无意于此的旁观者在为他事辛劳的间隙，也能够了解个中的美妙。

如果读者认为此书稍有所裨益，那么就烦请您到书店，或通过网络去购买此书。Linux 是一个有待去探索的深奥领域，我自信本书能为您提供一点线索，甚至作为随身指导，常常温故以知新。您的付出，无论相对于您的收获，还是相对于本书创作过程中的付出，都是值得的。当然，笔者也不避言，因为自己还是一个有志于其他山峰的攀登者，希望为下一次的旅程积累一些资本。

本书目标

大多数爱好者在阅读 Linux 内核源代码时会产生这样的困惑，我们很少能找到针对 Linux 操作系统，甚至某个单独的内核模块，在设计和开发方面的文档。仅有 Linux 社区的一些高手们对一些关键算法或者一些设计考虑的讨论。此外，当前大多数的源码分析书籍，都只是就函数或代码进行解释，没有给出整体和全面的视角。对于处于新手上路，或者小技初成级别的读者，只能获得局部和片面的认识，在理解这些讨论和阅读内核源码时会非常困难，常常产生挫败感。

写作本书的初衷，就是希望从软件设计和开发者的角度对与 Linux 存储相关模块作一个梳理。我们将设计一些主要的场景，跟踪实现的各个层次，对其中的主要函数进行代码级的讲解。在分析每个模块时，会给出整体框架与主要数据结构之间的关系，并列出各个域的详细含义，比起单调的代码阅读理解，相信会达到更好的效果。

本书在构思之初，所计划的目标远非上面列举的几项。笔者曾经试图加入设备仿真原理部分，分析虚拟机是如何仿真 CPU、PCI 子系统、SCSI 子系统、存储适配器、网络适配器以及磁盘设备的。笔者还曾设想从 Linux 内核开发者的角度，尽量贴近软件设计和开发文档，讲解各个子系统和模块的构思过程、设计思想、技术难点及解决方案等。不过由于时间和精力所限，最后有所取舍。

本书最终确定命名为《存储技术原理分析——基于 Linux2.6 内核源代码》。坦率地说，这只是最初设定的写作目标中基础的部分，此外，还准备了网络子系统、iSCSI、NFS、Heartbeat 虚拟机等诸多素材，分别被规划为进阶篇、高级篇的内容。至于最终会不会有机会呈现给读者，那就要看本书的市场反应和笔者的精力了，本书旨在通过分析 Linux 内核源代码讲解存储、网络和虚拟机的相关技术。如果读者有足够的耐心和毅力跟随我们完成这一次旅程，您将理解以下几点：

- 设备发现过程：了解操作系统如何发现 PCI 设备、SCSI 设备、块设备，并和驱动绑定起来；
- 存储 I/O 路径：了解用户对文件的读/写请求怎么通过 I/O 路径，最终到达磁盘介质上；
- 内核编程模式：理解 PCI-SCSI HBA 驱动、块设备驱动，以及文件系统等编程框架。

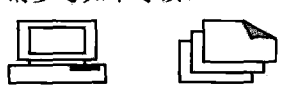
本书读者

本书所针对的读者应该具有一定基础。应该对存储技术有些了解，应该对操作系统有些领悟，应该对软件开发有些心得，但是这种了解、领悟和心得又似乎浮于表面，经不住考验。最为重要的是，应该有对细节的渴望和对未知的好奇。本书通过分析 Linux 源代码系统阐述存储技术的内在原理，它要告诉的，不是 Linux 的历史有多长、功能有多强，或是那些古老的话题，重复在我们的身旁；它要讲述的，是一些执着的人们思考着什么，又最终决定了什么，在并不为我们所知的地方。

本书导读

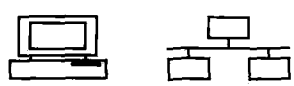
本书针对的 Linux 内核版本号为 2.6.34，并且以 x86 为硬件平台为目标。为了突出重点，将在各个部分讨论最具普遍性、而不是最新的技术。例如，会探讨 PCI 和 SCSI，但不会涉及 PCIe 或者 SAS。至于文件系统部分，还是以 Minix 作为基础，尽管当前有很多非常强大的文件系统存在，并且还会不断被推出。在行文过程中，可能会谈到某些厂商的产品或驱动，这绝非出于市场考虑，也并不表明该产品是行业内功能最强或性能最优的，或者这些驱动是“没有错误（bug free）”的。

本书涉及的内容较多，分为三部分共 8 章。为了使读者在阅读过程中不至于迷失，请参考如下导读。



1. 第一部分为第 1 章，介绍存储技术的概况。

这一章讨论存储系统的组成元素：磁盘驱动器、存储设备、服务器部件和存储软件，阐述和存储相关的各种技术，如备份、快照、连续数据保护、重复数据删除、虚拟化及虚拟机技术等。

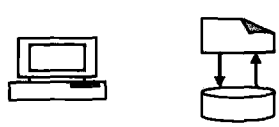


2. 第二部分包括第 2 章到第 4 章，介绍设备发现的过程。

第 2 章讲述 Linux 驱动模型。它为 Linux 内核构建了一个综合的、统一的机制，以内核对象为基础，将总线、设备和驱动关联起来，组织成一个层次结构的系统，方便了各种类型设备的热插拔和电源管理，同时借助 sysfs 提供了一个完全层次结构的用户视图。

第 3 章讲述 PCI 子系统。它将 SCSI 适配器、网络适配器等设备连接到主机 I/O 总线。

第 4 章讲述 SCSI 子系统。它将磁盘或者外部存储设备连接到主机 I/O 总线。



3. 第三部分包括第 5 章到第 8 章，介绍存储 I/O 的路径。

第 5 章讲述块 I/O 子系统。它向上层提供 I/O 请求 API，并实现 I/O 调度，将请求派发到具体块设备的请求

队列执行，以及提供请求完成的下半部处理 API，直至最终调用上层的请求完成回调函数结束 I/O。

第 6 章讲述 Linux RAID。MD 模块是一个虚拟块设备层。这一章将以 MD 模块源码为基础，阐述 Linux 内核中如何支持各种不同的 RAID 级别。

第 7 章讲述 Device Mapper。这一章将以 Device Mapper 模块源码为基础，阐述设备映射原理及各种映射规则的实现。

第 8 章讲述文件系统。文件系统是存储和组织文件（即一系列相关的数据），以便可以方便地进行查找和访问的一种机制。这一章将介绍 Linux 内核如何通过虚拟文件系统层，实现对各种具体文件系统的支持，以及从装载文件系统到访问文件等系统调用的流程。

本书图例

本书不会仅限于从概念上来讲解存储技术，而是以 Linux 内核源码为基础进行详细的阐述。理解 Linux 内核的关键在于把握各个数据结构之间的关系，用图示方式描画出各个数据结构之间的关系无疑是最为直观的。在描画各个关系图时，本书遵循如图所示的图例注解：

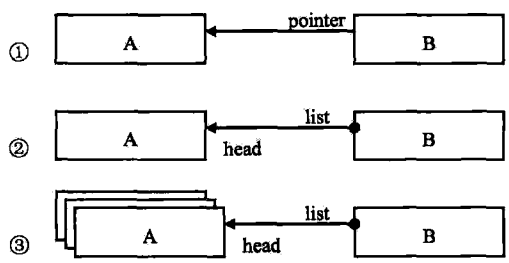


图　本书图例注解

图中①表示数据结构 B 有一个 pointer 域指向数据结构 A。它的示例代码如下：

```
struct B {
    ...
    struct A *pointer;
    ...
};
```

②表示数据结构 B 通过 list 域链入数据结构 A 的以 head 域为表头的链表，它的示例代码如下：

```
struct A {
    ...
    list_head head;
    ...
};
struct B {
    ...
    list_head list;
    ...
};
...
void link_B_to_A(struct A *a, struct B *b) {
    list_add(&b->list, &a->head);
}
```

我们看到，在数据结构 A 中定义了 head 域，数据结构 B 中定义了 list 域。两个域都是 list_head 类型，这是 Linux 内核代码中表示链表的数据结构，在分析代码中遇到它再进行分析。上面的例子中还有一个函数 link_B_to_A，用于将结构 A 的对象通过 list 域，链接到结构 B 的对象以 head 域为首的链表中，其中 list 域被称为连接件，head 域被称为表头。

③表示数据结构 B 通过 list 域链入哈希表的某个哈希桶中。哈希桶的数据结构为 A，其 head 域为链表表头。

表示结构 A 的数组。

感谢与联系方式

本书是一本技术性组织的书籍，能通过对 Linux 源码的分析来讲解存储、网络和虚拟机相关技术的内涵，外延及这些技术的原理。在写作过程中笔者查阅了很多资料，在此一并向有关资料的提供者表示感谢。

感谢我的导师上海交大计算机系的白英彩教授！白老师是我一生的导师，自毕业以来，始终给予我职业上的指导，并提供了许多宝贵的建议和机会。

感谢博文视点公司的几位编辑。如果不是有幸遇到他们，本书可能还只是孕育着的一个想法。在写作过程中，他们提供了很多的专业指导，尤为感动的是，在因延期交稿而忐忑时，还给予了从作者角度设身处地的劝慰，使得我安心、愉快地继续创作。

真诚地感谢我的家人，让我在紧张写作的间隙，总能够享受家的温馨。尤其是我的爱人邓玉洁，对于我的每一个想法，尽管可能不见得是主流，她总是给予极大的支持。由于她的努力，我可以无需过多顾虑生活上的压力，感谢她，还有我们生命中最重要的小末末！

最后，感谢我所有的朋友，以前的和现在的，有联系的和没有联系的。曾经为远离从前的朋友而遗憾，怕从此友情不再。现在终于释怀，人生就好像一个旅程，朋友就是我一路的风景，不同的阶段风景会不同，感谢他们，见证了我生命的历程。特别的感激送给：柴剑锋、陈颖、戴广成、丁旭华、董歆奕、方敏、谷大武、黄昉、蒋川群、蒋文蓉、金崇英、李虎、李静、李小勇、梁坚、刘俊、刘伟、任松林、田忠、王安保、王金浦、杨向群、张冠群。同样，向很多没有列出名字的朋友说声抱歉——你们在我心中。

当然，由于资料难觅，加之笔者水平有限，本书难免存在不少的瑕疵和错误的理解，敬请读者批评指正，可以发送邮件至 ao_qy@hotman.com 进行联系。

最后附上小诗一首，结束全文。

【登南岳赋】

二零一零年，清明，霖，游寿岳衡山，经半山亭，南天门等，至祝融峰晋香，感自然浩瀚，叹世俗虚华，以此诗记之：

半山烟雨半山空，一路流云一路松，
携风漫道南门外，相邀长住寿天宫。

敖青云
2010 年 7 月～2011 年 2 月
写作于上海浦东新区图书馆
完稿于上海第二工业大学

目　录

第一部分　存储技术

第三部分 存储 I/O

图索引