重剑无锋 大巧不工

大话存储

原理精解与最佳实践
网络存储系统

张冬 编著

★ 国内首次全面披露网络存储深层技术细节
★ 特立独行的行文风格，一针见血的诠释技术
★ 网络存储和存储网络，谈笑间难点灰飞烟灭
★ 全面解密SAN、NAS系统
★ 涉及众多底层细节，中文资料独家提供
★ 提供最佳操作实践，严禁纸上谈兵
★ 四大网站本书讨论专题，你不是一个人在战斗

清华大学出版社

大话存储——网络存储系统原理精解与最佳实践

张　冬　编著

清华大学出版社
北　京

内 容 简 介

网络存储，是近二十年来的新兴行业。从纸带到硬盘再到大型磁盘阵列，存储系统经历了从简单到复杂，从单块硬盘到存储区域网络(SAN)。网络存储行业目前已经是一个步入正轨的IT行业了。

网络存储是一个涉及计算机硬件以及网络协议/技术、操作系统以及专业软件等各方面综合知识的领域。目前国内阐述网络存储的书籍少之又少，大部分是国外作品，对存储系统底层细节的描述不够深入，加之术语太多，初学者很难真正理解网络存储的精髓。

本书以特立独行的行文风格向读者阐述了整个网络存储系统。从硬盘到应用程序，这条路径上的每个节点，作者都进行了阐述。书中内容涉及：计算机IO基本概念，硬盘物理结构、盘片数据结构和工作原理，七种常见RAID原理详析以及性能细节对比，虚拟磁盘、卷和文件系统原理，磁盘阵列系统，OSI模型，FC协议，众多磁盘阵列架构，SAN和NAS系统，TCP和以太网以及IP SAN，协议融合理论，存储虚拟化，存储及服务器群集，数据保护和备份技术，快照技术，数据容灾技术。

本书用独特的写作方式通俗地诠释了这些晦涩、枯燥的难点技术并提供了许多前所未有的操作实践和本书作者长期从事存储工作的一些经验点滴。

本书适合初入存储行业的技术工程师、售前工程师和销售人员阅读，同时适合资深存储行业人士用以提高技能，另外，网络工程师、网管、服务器软硬件开发与销售人员、Web开发者、数据库开发者以及相关专业师生等也非常适合阅读本书。

图书在版编目(CIP)数据

大话存储——网络存储系统原理精解与最佳实践/张冬编著. —北京：清华大学出版社，2008.11
ISBN 978-7-302-18672-4

Ⅰ. 大…　Ⅱ. 张…　Ⅲ. 计算机网络—信息存储—研究　Ⅳ. TP393.0

中国版本图书馆CIP数据核字(2008)第150498号

责任编辑：栾大成
封面设计：杨玉兰
版式设计：北京东方人华科技有限公司
责任校对：李玉萍
责任印制：王秀菊
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦A座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：清华大学印刷厂
装 订 者：北京鑫海金澳胶印有限公司
经　　销：全国新华书店
开　　本：185×260　**印　张**：28.5　**字　数**：684千字
版　　次：2008年11月第1版　　**印　　次**：2008年12月第2次印刷
印　　数：5001～8000
定　　价：58.00元

前　言

各位朋友，大家好。感谢您购买并阅读此书。我叫张冬，网名冬瓜头，山东人。承蒙各位朋友的大力帮助，让我写成了这本书。

创作背景

写这本书的初衷，就是为了让广大系统工程师、IT 工作者对网络存储系统能有一个深入的了解。市面上已经有一些讲述网络存储系统的书籍，但是我发现对于初学者来说，这些书籍大多太过抽象晦涩，语言不够通俗易懂。而对于专业的网络存储工程师来说，这些书又不是必读之物。所以我就产生了写一本任何 IT 人都能看懂并且津津乐道的关于网络存储方面的书的想法。

创作过程

然而，当我真正新建一个文档准备往里敲字的时候，却发现，写作不是那么容易的，尤其是写一本让人人都能看懂的书。本书开稿日期我记得是在 2006 年 9 月。开始的时候，在 3 个月时间内就已经写完了大体的框架。然而，这个框架内容应该说是想到什么就写什么，也就是把我脑海里所有知道的东西都写出来，非常凌乱，更别说语言措辞了。然而，爆发的快，熄灭的也快。随后的几个月内，我再也找不到灵感应该写什么，怎么写了。这完全是由于本人知识匮乏所导致的。

第一阶段写作的过程中，随着写作进行，遇到的理论问题也越来越多。我晚上进行写作和思考遇到的问题，白天就上网查阅这些问题的理论根基，对于一些没有答案的问题，我就自己思考，琢磨答案，求助周围和网络上的高手们。

框架写成后，就是漫漫的充实之路。每天晚上，有了灵感，就修改一下，续写一下，没有灵感，就闭目欣赏音乐。

就这样，时间一直走到了 2007 年 6 月。我在网络上遇到了清华大学出版社编辑栾大成。这位编辑的随和与热心，深深地打动了我。和他交谈时，我感觉他是在帮助我写书，而不是我在求他为我出书。从此，我坚定了要把这本书写完的决心。我重新打开未完成的文档，业余时间全部投入到写作中。就这样，又写了 8 个月。这 8 个月中，大部分时间是在整理写完的章节，包括修改技术错误以及语言方面的不当。修改比新写要困难许多，有时候遇到一些不合适的地方，甚至需要整章都要重新构思并写作。

这本书的写作可以说是跨越了三地，先后为北京(2006 年 9 月开稿于北京西二旗燕尚园)、青岛、大连，最后于 2008 年 2 月在大连收稿，接下来就是漫长的修订、完善过程。

读者对象

本书适合于已经或者打算进入存储领域的 IT 工程师阅读，同时也适合所有 IT 工作者和具有一定计算机系统知识和网络知识的读者。本书既适合初入存储行业的技术工程师、售前工程师和销售人员阅读，又适合资深存储行业人士以便提出宝贵意见和建议。

内容提要

第 1 章　盘古开天——存储系统的前世今生

介绍存储的历史和现状，各种需要掌握的主流技术。

第 2 章　IO 大法——走进计算机 IO 世界

本章介绍计算机系统是如何进行 IO 动作的。阐释了 CPU、内存、磁盘三者是怎么联系起来并且从磁盘读写数据的。给出计算机总线的概念，CPU、内存和磁盘三者都在总线上，通过协议来互相交换数据。这一章是作为读者继续理解存储系统的基础。

第 3 章　磁盘大挪移——磁盘原理和技术详解

本章介绍磁盘，包括磁盘的构造、原理、如何寻址、外部接口、高级磁盘技术。其中用了众多的类比，通俗讲述磁盘从存储数据、传输数据的详细过程。

第 4 章　七星北斗——大话/详解七种 RAID

本章是读者真正进入存储子系统的入口，讲述了 RAID 技术。前面部分使读者感性认识了 RAID 技术，后面部分在有了轮廓之后，从纯技术角度更深入地解释了 RAID 技术的细节。

第 5 章　降龙传说——RAID、虚拟磁盘、卷和文件系统实战

本章前面部分描述了在 RAID 技术基础上的一些更加高级的存储技术，包括 RAID 控制器、RAID 卡、虚拟磁盘，虚拟卷、卷管理。每个部分又分成多个细节部分来阐释。后面部分则描述了存储系统的一个重要层次，即文件系统。用了一个仓库模型来阐释了文件系统的作用原理。

第 6 章　阵列之行——大话磁盘阵列

本章是承前启后的一章，描述了磁盘阵列的发展，从最简单的 JBOD 模式的磁盘阵列开始，一直到高端复杂的带有 RAID 控制器的大型磁盘阵列。最后达到本章的高潮，即随着存储系统向主机外部扩展，其最终形态是网络存储系统，引出了本书后面部分要阐释的主体——存储区域网络(SAN)

第 7 章　熟读宝典——系统与系统之间的语言 OSI

既然存储系统已经达到了网络化的程度，那么非常有必要向读者阐述一下网络 OSI 模型。本章用了各种比喻，向读者展现了 OSI 模型中的 7 个层次。

第 8 章 勇破难关——Fibre Channel 协议详解

本章阐释了 Fibre Channel，一种广泛用于后端存储网络的网络技术。用以太网和 Fibre Channel 网络做了比较，并且虚拟了一个钻研存储技术的人物角色来完成这一章。这个角色根据 OSI 模型和以太网的实现思想，创造了 Fibre Channel 各个层次的协议系统，并以其得天独厚的优势成为了最适合存储网络适用的协议。

第 9 章　天翻地覆——FC 协议的巨大力量

本章写了 Fibre Channel 协议用于存储网络之后，对传统的基于 SCSI 的磁盘阵列架构带来的天翻地覆的变化。Fabric 协议将传统的磁盘阵列的前端和后端协议统统替代了，实现了存储系统的彻彻底底的网络化改造。

第 10 章　三足鼎立——DAS、SAN 和 NAS

本章阐释了现今存储领域的三大主要架构，即 SAN、NAS、DAS。讲述了 NAS 的由来。Fabric 将存储系统彻底网络化，而各种五花八门的协议又使文件系统也被网络化了。当今世界，任何东西都和网络扯上了关系，甚至洗衣机，冰箱都做上了以太网接口。没有什么不可以网络化的，只要有通信的需求，就可以网络化。

第 11 章　大师之作——大话以太网和 TCP/IP 协议

本章是对下一章将要阐释的 IP SAN 所做的一个铺垫。要理解新兴的 IP SAN，必须对 TCP/IP 协议有一定理解。本章向读者展现了 TCP/IP 协议的设计思想和实现方式，对以太网也做了说明。

第 12 章　异军突起——存储网络的新军 IPSAN

本章阐释了 IP SAN 这个存储网络领域的新军。Fabric 可以用于存储网络的协议，IP 也可以。它们都是用于网络互联的协议，IP 要在存储网络分一杯羹，且看 IP 有何过人之处可以让其和 Fabric 协议一决高下呢？本章将给出答案。

第 13 章　握手言和——IP 与 FC 融合的结果

夫天下之势，分久必合，合久必分。网络存储领域也遵循这个规则。IP san 和 Fabric san 激烈的竞争，表面上使其二者互相排斥远离，但还是那句话："本是同根生，相煎何太急？"。没错，二者各有长处，也各有短处，为何二者不能合作，互相取长补短，形成新的协议体系呢？完全可以。本章就描述了这样两种新的协议体系，IP 和 Fabirc 协议互相融合。作者在本章引入了一个新概念，即"通信协议间的相互作用"，并对这个概念做了深刻的比喻和透彻的阐释。

本章是本书的高潮，纵观本书，从一开始向读者介绍计算机总线、磁盘，到后来逐渐将磁盘向外扩充，形成盘阵，然后将盘阵与主机的连接网络化，然后将盘阵自身的各个模块彻底网络化。网络化之后，就是各种网络协议用于这个网络，再后来，各个协议之间相互融合，达到最高境界。

第 14 章　变幻莫测——虚拟化

虚拟化这个词，在计算机系统中无处不在。本章从物理层一直到应用层，向读者阐释了虚拟化在计算机系统各个层次中的作用原理。

第 15 章　众志成城——存储集群

随着小型机和 PC 的成本不断降低，以前需要大型机方能进行的运算，现在也可以运用小型机甚至 PC 组成的集群系统来进行。本章向读者阐述三种集群(HA 集群、LB 集群、HPC 集群)的概念和作用原理，以及集群系统中的存储子系统的一些概念和特点，包括集群文件系统等知识。

第 16 章　未雨绸缪——数据保护和备份技术

本章讲述了与存储密切相关的一个领域，就是数据保护和备份领域。面对庞大的数据，如何保证其安全性？在这一章，将向读者展现数据备份的方方面面。

第 17 章　愚公移山——存储容灾技术

本章描述了容灾系统的各个组成要件，从一个通俗的例子，一步一步带领读者认识到容灾系统的精髓思想。

附录　五百年后

最近，固态硬盘和芯片存储技术炒得火热。我本人也相信，不超过 10 年或者更短时间，机械硬盘将彻底被逐出市场。故障率高、费电量大、体积庞大、速度已经达到技术极限等这些缺点似乎已经决定了机械硬盘的命运。而存储介质被替换成芯片之后，整个系统的架构就可能发生革命性的改变。本章中，作者对未来的系统架构做了自己的预测。

阅读指南

读者根据自己的理解程度和水平，可以略过一些认为已经掌握的内容。但是推荐读者顺序阅读每一章，以防由于缺乏连贯性导致的对后续章节的某些细节造成的误解。

致谢

感谢中国民航信息网络股份有限公司研发中心的曹迎军、张博、丁玎、乔靖四位同志的帮助！他们在很多计算机架构以及操作系统底层技术方面给了我大力支持，曹兄还在方法论方面深深的影响了我。同时也感谢中国航信研发中心的高新同志！

感谢存储在线论坛以及 Cisco 网络技术论坛的各位网友。如果没有你们的热烈参与就没有这本书的面世！

感谢清华大学出版社的栾大成编辑以及其他参与本书出版的工作人员！你们的热心帮助，才使得这本书从写成到出版一气呵成！

最后，感谢我的父母、女友。爱你们到永远！感谢母亲的谆谆教诲，为了儿子操劳了一辈子。感谢女友帮我打点生活。

作者联系方式：冬瓜头
Email：myprotein@sina.com
QQ：122567712
MSN：myprotein0007@hotmail.com
BLOG：HTTP://space.doit.com.cn/35700

作者水平有限，计算机技术无限，书中错误在所难免，希望广大读者纠正。谢谢！

声明

1. 本书部分图片和内容来自于互联网和相关著作，版权归原作者所有，本书引用目的只是为了辅助读者对本书主题的理解。

2. 本书对某些产品的分析引用了相关产品的官方图片用以辅助说明主题，版权归原厂商所有。

3. 本书中所介绍的产品不带有任何主观偏向色彩，所使用的产品相关图片没有任何主观不良意图。如有错误之处请提出，将在下一版中改正。谢谢！

编　者

编 辑 序

第一次接触张冬，是在 2007 年 8 月，张冬 QQ 加我，说有本书问我感不感兴趣。所有编辑可能都对送上门的没有经手策划的书主观的轻视，况且是这样一个“非主流”的选题。经过长时间的沟通，我发现这是个真诚且严谨的家伙，同时在论坛中我发现张冬的作品负面评价很少，而且人气很高。记得无论我什么时候上线，他总在，这又是一个十分努力的家伙。另外，我了解到，张冬在 IT 行业算是半路出家，他是化学专业出身，但就是这样一个跨行业的人，却能用清晰的文笔来描述网络存储这样相对晦涩的技术！

一个真诚、严谨且努力的技术高手……这样的人的作品怎么会不好呢？

一个脱离技术多年的策划编辑凭什么为一本技术性很强的 IT 图书作序？汗颜……我以前也研究过“存储”，仅仅局限于硬盘结构，10 年前写的《实战 DEBUG》、《汇编语言超浓缩教程》系列文章涉及到对硬盘的分析和操作，这些文章现在在网上还能找到，也经常有朋友或作者跟我聊起来，得意洋洋……曾经立志成为存储达人，然而天赋有限，未遂。

张冬的作品从收到稿件阅读第一章开始，我就感觉这一定是本好书。是当年梦寐以求的资源。难得的是，这本书的行文异乎寻常的流畅，以致我曾经问过张冬：“小样儿，你是学中文的吧？”一本专业性极强的图书，最关键的就是要把问题讲清楚。

张冬用一种“另类”的方式对一些晦涩的概念和理论进行了重新包装。充斥着“庸俗的”解释与描述，比如：数据包在网络中的流动过程——是对照快递公司的业务流程比对讲解的，容易理解而且印象深刻。

另外，这本书提供了一些培训级别的操作。大家知道类似网络存储这种规模的部分操作，很少能在家里用 PC 来进行实际操作(当然模拟练习还是可以的)，张冬有条件在这样的专业操作环境进行操作步骤的整理，这些细致的重量级操作也是本书另外的价值所在。

信息存储是这个世界的未来，将来我们的一举一动的背后都会伴随大量的信息存储行为，存储已经成为了一个行业，任何动作都离不开它。现在，网络工程师，网管，Web 开发者，数据库开发者，软件开发(特别是网络应用)者都必须掌握网络存储的一些细节，可以说基本上所有 IT 技术从业者都需要或多或少地了解存储。这一定会是个广大的市场，或者说已经是广大市场了。

这样的一本书，我希望并且相信会给大家的学习带来帮助，也相信这样一本特立独行的好书能够让大家很多年以后还能回忆起来并津津乐道地向朋友推荐。

清华大学出版社　栾大成

目　录

第一章

盘古开天

存储系统的前世今生

- 存储历史
- 存储技术

数据存储是人类千百年来都在应用并且探索的主题。在原始社会，人类用树枝和石头来记录数据。后来，人类创造了铁器，用铁器在石头上刻画一些象形文字来记录数据。而此时，语言还没有形成，人们记录的东西只有自己才可以看懂。

随着人类相互之间交流的愿望越来越迫切，逐渐形成了通用的象形文字。有了文字之后，人们对每个文字加上了声音的表达，就形成了语言，也就是将一种形式的信息，转换成另一种形式的信息。人们用文字作为交流工具，将自己大脑产生的信息，通过这种方式传递给其他人。这和网络通信的模型是一样的，计算机将数据利用TCP/IP协议，先通过网卡编码，再在线缆上传输，最终到达目的地。人类将大脑中的数据，变成语言编码，然后通过嗓子的振动，通过空气这个大广播网，传递给网内的每个人。

后来，人们将文字刻在竹片上保存。再后来，蔡伦发明了造纸技术，使得人们可以将信息写到纸上，纸张摞起来就形成了书本。后来，毕昇用泥活字革新了印刷术，开始了书本的印刷。再后来，激光打印取代了活字板。再后来，纸带、软盘、硬盘、光盘等方式出现了。再往后，就需要广大科学工作者去努力发明新的存储技术了。

1.1 存储历史

存储在这里的含义为信息记录，是伴随人类活动出现的技术。

1. 竹简和纸张

竹简是中国古代使用的记录文字的工具，后来被纸张所取代，如图 1.1 所示。

2. 选数管

选数管是 20 世纪中期出现的电子存储装置，是一种由直观存储转为机器存储的装置。其实在 19 世纪出现的穿孔纸带存储就是一种由直观存储转向机器存储的产物，它对 19 世纪西方某国的人口普查起到了关键的加速作用。

选数管的容量从 256～4096 比特不等，其中 4096 比特的选数管有 10 英寸长，3 英寸宽，最初是 1946 年开发的，因为成本太高，并没有获得广泛使用。图 1.2 是容量为 1024 比特的选数管。

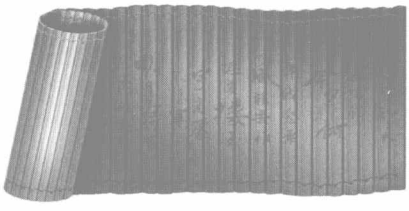

图1.1 竹简

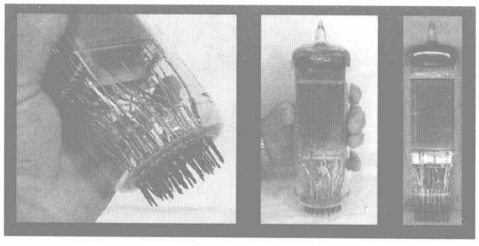

图1.2 选数管

3. 穿孔卡

穿孔卡片用于输入数据和程序，直到 20 世纪 70 年代中期仍有广泛应用。图 1.3 和图 1.4 是一条 Fortran 程序表达式 Z(1) = Y + W(1)所对应的穿孔卡和穿孔卡片阅读器。

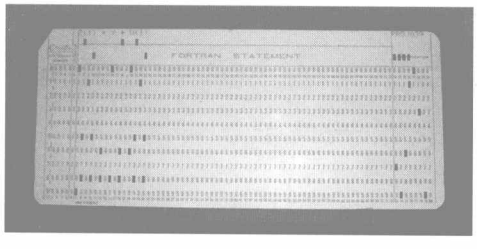

图1.3 穿孔卡

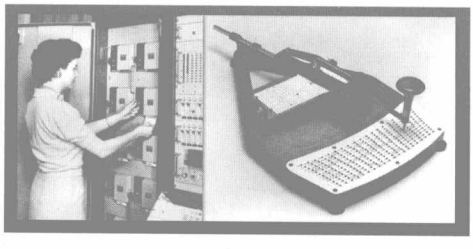

图1.4 穿孔卡片阅读器

4. 穿孔纸带

穿孔纸带用来输入数据，输出同样也是在穿孔纸带上。它的每一行代表一个字符，如图 1.5 所示。

5. 磁带

磁带是从 1951 年起被作为数据存储设备使用的。磁带在当时被称为 UNISERVO。图 1.6

所示的最早的磁带机可以每秒钟传输 7200 个字符。如图 1.6 所示这套磁带长达 365 米。

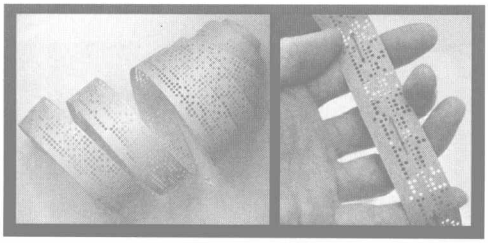

图1.5 穿孔纸带

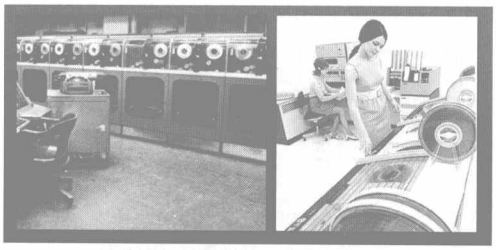

图1.6 磁带及磁带机

从 20 世纪 70 年代后期到 20 世纪 80 年代出现了小型的盒式磁带，长度为 90 分钟的磁带每一面可以记录大约 660KB 的数据，如图 1.7 所示。

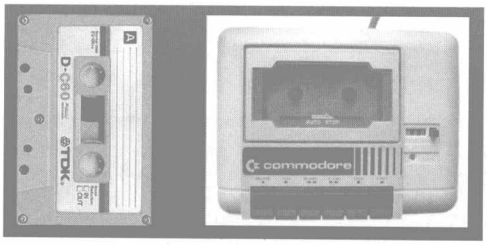

图1.7 小型盒式磁带

6. 磁鼓存储器

磁鼓存储器最初于 1932 年在奥地利被创造出来，在 20 世纪五六十年代被广泛使用，通常作为内存，容量大约 10KB，如图 1.8 所示。

7. 硬盘驱动器

第一款硬盘驱动器是 IBM Model 350 Disk File，如图 1.9 所示，于 1956 年制造，其中包含了 50 张 24 英寸盘片，而总容量不到 5MB。

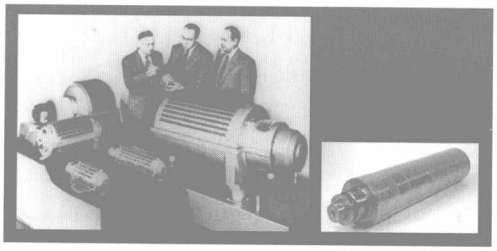

图1.8 磁鼓存储器

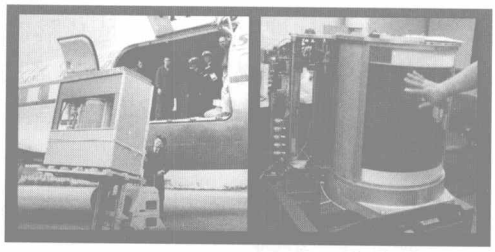

图1.9 早期的硬盘驱动器

首个容量突破 1GB 的硬盘是 IBM 在 1980 年制造的 IBM 3380，如图 1.10 所示，总容量为 2.52GB，重约 250 千克。

8. 软盘

软盘由 IBM 在 1971 年引入，从 20 世纪 70 年代中期到 20 世纪 90 年代末期被广泛使用，最初为 8 英寸盘，之后有了 5.25 英寸和 3.5 英寸盘。1971 年最早的软盘容量为 79.7KB，