Design and Optimization of Erasure-coded Clustered Storage Systems

纠删码存储集群系统设计与优化

Huang Jianzhong, Cao Qiang
Qin Xiao, Xie Changsheng

黄建忠 曹 强 秦 啸 谢长生 著

科学出版社

纠删码存储集群系统设计与优化

Design and Optimization of Erasure-coded Clustered Storage Systems

黄建忠　曹　强　秦　啸　谢长生　著

Huang Jianzhong, Cao Qiang, Qin Xiao, Xie Changsheng

科学出版社

北　京

内 容 简 介

本书以纠删码存储集群为研究对象，分别阐述编码原理和存储机制，对正常用户读、正常用户写、失效用户读、离线重构、在线重构、节点扩容、副本归档和存储节能等功能需求展开专题讨论，每个专题在分析现有研究的基础上，给出总体研究思路、具体设计方案和对比验证实验，可为分布式存储系统设计提供方案借鉴和技术参考。

本书适合计算机等相关专业的高年级本科生和研究生阅读，也可作为从事存储系统设计与管理的工程技术人员的参考书。

图书在版编目 (CIP) 数据

纠删码存储集群系统设计与优化 / 黄建忠等著. —北京：科学出版社，2016.4

ISBN 978-7-03-047577-0

Ⅰ. ①纠… Ⅱ. ①黄… Ⅲ. ①信息存贮－系统设计 Ⅳ. ①TP333

中国版本图书馆 CIP 数据核字 (2016) 第 046609 号

责任编辑：余 丁 赵艳春 / 责任校对：桂伟利

责任印制：张 倩 / 封面设计：迷底书装

科 学 出 版 社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

文林印务有限公司 印刷

科学出版社发行 各地新华书店经销

*

2016 年 4 月第 一 版 开本：720×1 000 1/16

2016 年 4 月第一次印刷 印张：12 3/4

字数：243 000

定价：69.00 元

（如有印装质量问题，我社负责调换）

前　言

随着企业、机构、个人对信息技术的依赖程度提高，越来越多的用户数据和业务数据被存储在计算机系统中。数据信息化在提高工作效率和业务水平的同时，也带来了数据丢失的风险，一旦系统中关键数据丢失或损毁，将导致不可估量的直接和间接经济损失，因此需要提高存储系统可靠性以增强数据可用性。为了保障存储可靠性与数据可用性，必须采用数据冗余机制。冗余机制主要包括副本和纠删码技术，前者将每个原始数据分块都镜像复制到另一个存储设备上，以保证原始数据不可用时有副本可恢复；后者将原始数据进行划分、编码和分发，最后将编码分块独立存储到存储设备上。相对于副本技术，纠删码具有更高的存储效率（即存储空间利用率），在相同存储空间情况下能获得更高的容错能力。如今，纠删码已成为副本之外的另一种可靠存储方式，广泛应用于存储系统中，如面向大数据挖掘的集群存储、面向数据托管的云存储和面向数据长期保存的归档存储等。

本书以纠删码存储集群系统为研究对象，针对其设计和优化展开论述。相对于磁盘阵列仅涉及异或运算和磁盘读写，纠删码存储集群需要考虑线性组合运算、网络传输、磁盘读写三个方面，其系统结构、I/O 访问路径及性能衡量指标都有所不同，因此，怎样设计纠删码存储集群系统以提高用户访问性能、实现高效率存储维护是一项具有挑战性的研究课题。本书整体分为两大部分：一部分，从系统的高度梳理并归纳了纠删码存储集群所涉及的纠删编码和数据存储两方面研究，既可以从编解码计算、网络传输和 I/O 访问三个环节来研究纠删码集群支撑技术，又可以从编码/布局和存储/调度两个层面来设计并优化纠删码集群；另一部分，结合纠删码存储集群所面临的前台用户访问 I/O 和后台系统维护 I/O，分别对正常用户读、正常用户写、失效用户读、离线重构、在线重构、节点扩容、副本归档和存储节能等八种实际功能需求展开专题讨论，在每一个研究专题中给出具体的设计方案和实际的优化案例。

分布式系统受到一致性、可用性和分区容错性原则（Consistency-Availability-Partition Tolerance，CAP）的约束，在可靠存储系统中，系统架构师难以提出同时满足一致性、可用性和分区容错性的设计方案。纠删码存储集群的设计初衷是后两者，即保障系统可用性和容错性，从而只能降低数据一致性要求，鉴于此，纠删码存储集群的一个主要应用场景是具有写一次读多次（Write-Once-Read-Multiple，WORM）特性的 I/O 场景，如监控视频的集中存储、灾备数据的远程归档等。相似地，在为纠删码存储集群系统设计各种方案时（如重构、归档、扩容、节能等），感触最深的是，很难提出十全十美的存储系统方案，某一项性能指标的提升往往以牺牲另一项性能指标为代价，例如，面向三副本的纠删码归档方案中，流水线方式能增强归档并行性，但

需要三倍的磁盘读开销；存储节能方案中，采用缓存写能推迟休眠节点唤醒时间以提高能效性，但需要冒着缓存数据丢失的风险。本书通过八个具体研究案例剖析了纠删码存储集群系统的技术难点，并有针对性地提出具体设计思路，特别是以指标权衡为导向，深入讨论了各种优化策略的利与弊。

对于编码领域的理论工作者，纠删编码是一套基本的编码范式，而对于未接触过编码理论的存储系统架构师，纠删编码就显得有些神秘；相似地，对于存储系统架构师，存储访问和 I/O 调度是一组必备的方案技能，但对于未从事存储方案设计的编码理论工作者，对纠删编码在存储系统中的部署应用就存在一些疑惑。本书使用八个具体研究案例来分别缩小编码理论工作者和存储系统架构师对存储调度和编码布局的认知差距。所有研究案例都按“分析现有研究方案的优缺点 ⟶ 概括该研究主题的解决思路 ⟶ 提出针对性的设计方案 ⟶ 进行对比实验验证 ⟶ 剖析所提方案的适用性”进行层进式阐述。本书面向的读者包括计算机系统结构、信息通信工程专业的教师和研究生，以及从事分布式存储系统设计与管理的从业者（包括系统架构师、后台维护人员、编程人员等）。考虑到存储领域，特别是纠删码存储集群领域，国内外都缺乏能够理论结合实际的资料，相信本书能让广大读者加深理解纠删码存储集群的运行原理和访问机制，并能为分布式存储和存储可靠性研究、开发和产业发展提供一定帮助。

本书第 1～2 章阐述了纠删码存储集群的研究概况、研究主题、设计方案及编码相关技术；第 3～10 章给出纠删码存储集群各研究专题的具体方案。黄建忠、曹强、秦啸、谢长生参与了所有研究专题的方案讨论以及全书的统稿。研究生谢平、张峰豪、黄思倜、梁先海、罗海兵、王艳群、代尔卫参加上述研究方案的设计与实现，在此一并表示感谢。成书过程中得到了华中科技大学武汉光电国家实验室领导和师生的热心支持，书中使用了所在课题组的大量资料，在此致以诚挚的谢意。另外，本书各研究专题得到国家自然科学基金面上项目（编号：61572209）、国家 973 项目（编号：2011CB302303）、国家 863 项目（编号：2013AA013203）和武汉光电国家实验室创新基金的资助。

由于作者水平有限，书中难免存在不足之处，敬请读者批评指正。

作　者

2016 年 1 月于武汉

目　录

第1章 绪 论

“不要把所有鸡蛋放在同一个篮子里”——西方谚语

1.1 纠删码存储的研究意义

社会信息化程度不断提高，推动存储系统不断增加器件容量和数量。不幸的是，随着存储介质密度和存储器件数量的增加，器件及内部存储单元失效已成为经常性事件。考虑到数据蕴涵价值不断凸显，提高存储可靠性以保障数据可用性成为重要研究课题。纠删编码具有高存储效率和高容错能力，在大规模存储中具有突出优势，已广泛应用于国内外云存储系统中，但相比于副本机制存在编解码开销，又由于其特有 I/O 访问路径，纠删码存储存在诸多可改进的方面。

1.1.1 大规模可靠存储的应用需求

1. 数据量快速增长催生大规模存储

构建大规模存储系统的重要驱动力是应对呈爆炸式增长态势的数据量，尤其是为成千上万用户提高数据访问的数据中心。随着信息化时代的来临，人类正处于一个数据爆炸的时代，很多领域和行业需要应对呈爆炸式增长态势的海量数据。德国著名统计公司 Statista（www.statista.com）一份关于全球互联网数据量统计的报告显示，全世界的网民在 1998 年平均每月使用的网络流量为 1MB，2000 年为 10MB，2003 年为 100MB，2008 年为 1GB（1024MB 等于 1GB）[①]，而 2014 年平均每天使用的流量就达到 10GB。2001 年全球网络流量总和达到 10 亿 GB 的时间是 356 天，2004 年是 30 天，2007 年是 7 天，而在 2013 年只需 1 天，即一天产生的信息量就可以刻满 1.88 亿张 DVD 光盘。一个视频码率为 8Mbit/s 的摄像头一小时会产生 3.6GB 数据，一个城市若安装几十万个交通和安防摄像头，每月产生的数据量将达到几十 PB[1]。简言之，数据量快速增长已成为一种趋势，许多行业都面临数据快速增长这一挑战。

国际数据公司 IDC（www.idc.com）在 2012 年 12 月份发布了一份关于数字宇宙（digital universe）的研究报告，预计到 2020 年，数字宇宙即全球的数据量总和将达到 40ZB，相比 2005 年的 130EB，增幅达到 300 倍[2]。

① 1KB=1024Byte, 1MB=1024KB, 1GB=1024MB, 1TB=1024GB, 1PB=1024TB, 1EB=1024PB, 1ZB=1024EB

大数据浪潮也波及许多互联网企业，中国互联网有几个巨头公司，包括俗称 BAT 的百度（www.baidu.com）、阿里（www.taobao.com 或 www.tmall.com）和腾讯（www.qq.com）。对于以搜索引擎为主业的百度公司，其需要管理的搜索数据集已达到 EB 级，所存储网页数量以万亿页计算，每天超过 60 亿次搜索请求被处理，单日产生接近 PB 级的数据且其存储的网页总数量超过 1 万亿页，达到 EB 级别数据量；作为国内最大的电子商务平台，淘宝和天猫的注册会员超过 4 亿，年交易额突破万亿（2005 年 11 月 11 日，天猫交易额超过 912 亿），每天的活跃数据量已经超过 50TB，每天超过 6000 万人次访问；作为中国著名的互联网综合服务提供商，腾讯公司同样面临海量数据，根据 2013 年的统计数据，其通信应用 QQ 用户数达到 8 亿，存储数据总量经过压缩处理也达到 100PB，每天新增数据达到 200TB[3]。对于互联网企业，保存并管理好用户数据（账户信息、交易信息、家庭照片等）或为用户提供服务的后台数据（如搜索引擎的网页数据）是非常重要而必要的。为了存放海量数据，只能部署大规模存储系统；而随着大数据时代的来临，要求大规模分布式存储系统能不断增添新存储设备或升级旧存储设备来满足不断增长的存储容量要求。

2. 大规模存储面临存储可靠性挑战

随着计算机的网络化和全球化，越来越多的信息交流活动和工作事务被转移到网络上来，如今世界上几乎每一个国家都高度依赖通信、运输以及网络，包括政府事务、国防、金融、工商业等社会活动的方方面面。随着人们在日常生活和工作中对于信息技术的依赖程度提高，越来越多的重要数据被存储在计算机系统中，而信息系统的自动化也提高了业务的效率。对于公司、企业、机构甚至个人，其对信息系统的依赖程度也在逐步增加，同时也给用户带来了风险，一旦信息系统中的关键数据丢失或破坏，可能会给其带来不可估量的损失。假设存放在淘宝网站或腾讯网站上的用户数据（如身份证、银行卡、家庭照片、交易记录等隐私信息）发生大面积丢失或泄露，后果将不可想象，不仅用户会弃之而去，企业也将信誉受损、官司缠身。

根据 IDC 咨询公司的调查结果，美国在 20 世纪最后 10 年发生过数据灾难的公司中，55%当即倒闭，剩下 45%中由于信息数据丢失，29%在两年内倒闭，能生存下来的仅占 16%。而著名 IT 咨询服务提供商 Gartner 的报告数据也表明，40%的企业不能在灾难发生后恢复运营，剩下 60%中有 33%在两年内倒闭。美国 9·11 事件一年后，重返世贸大厦的企业由原先的 350 家减少到 150 家，200 家企业由于重要信息系统破坏及关键数据丢失而永远倒闭消失。据美国明尼苏达大学的一项研究报告显示，在灾害之后，如果无法在 14 天内恢复资讯作业，有 75%的公司业务会完全停顿，43%再也无法重新开业，因而有 20%的企业在两年之内被迫宣告破产。另外，对于灾害冲击的分析显示，各行业最长可忍受的信息系统停机时间分别为：金融业 2 天、销售业 3.3 天、制造业 4.9 天、保险业 5.6 天。平均看来，一般行业可忍受的信息系统停机时间为 4.8 天。如果以营业收入的损失来看，则金融业所遭受的损失最严重，可高达每日营业收入的

50%[4]。市场战略研究公司 Strategic Research Corp.（www.sresearch.com）的研究报告显示，ATM 系统中断 1 小时，平均损失为 1.45 万美元；银行业的信息系统中断，平均每小时的损失将达到 8.4 万美元；证券业的业务每停顿 1 小时，平均损失高达 650 万美元。

3. 数据冗余是存储可靠性保障手段

数据如此重要，作为数据存储的载体，存储系统的可靠性得到了极大的重视。就单个存储系统而言，其可靠性、I/O 延迟、访问并发数等性能指标日益提高，如 EMC、NetApp、IBM 等公司的高端存储系统，已被广泛应用于企业的核心业务系统。然而，一方面，由于数据的爆发式增长，存储规模不断扩大，每个单独的存储设备失效都可能影响整个存储系统，从而导致数据失效，如自然灾害、黑客攻击、病毒破坏、电力中断等；另一方面，高端存储系统意味着高昂的存储成本。一种行之有效的应对方式是将多个普通商业存储设备组合在一起，构建一个分布式存储系统。在构建分布式存储系统时，架构师需要处理多方面挑战：首先是数据剧增导致存储容量和性能的可扩展性需求；其次是服务连续性的保障，有的网络数据甚至要保证全年 365 天处于可访问状态；最后数据多样化、分散化以及非结构化的特点使得数据存储管理更为复杂。

如上所述，分布式存储系统要应对不断增长的容量和性能需求，而规模不断增长的存储系统需要考虑存储可靠性和数据可用性挑战，为了保证存储可靠性与数据可用性，常用的方法就是冗余，传统的冗余机制主要包括副本（replication）方式和纠删编码（erasure codes）。副本是将每个原始数据分块都镜像复制到另一个设备上，以保证原始数据失效时有副本可恢复；纠删编码最早起源于通信传输领域，通过对传输信息进行变换再进行传输，以解决数据传输中丢失/损耗这一问题。由于其具有防止数据丢失的特性，研究者将纠删码引入存储领域。如今，纠删码已经广泛应用于存储系统中，并取得了理想的效果。副本方式不涉及数据变换，而纠删编码会对数据信息进行变换和运算，得到支持数据冗余的编码数据，以（$k+r$, k）纠删码为例，其将一个原始数据分为 k 个数据分块，然后将其编码成 $k+r$ 个编码分块，并将编码分块分布存储在多个存储设备上。

1.1.2 副本方式具有低空间利用率

伴随着数据量的日益增大，存储系统规模也随之增大，包含的存储设备也越来越多，进而设备故障也随之增加。根据 Google 绿色能源团队项目经理 Jacobowitz 提供的耗电报告推测，截至 2011 年，Google 在全球的数据中心运行超过 90 万台服务器[5]。为了节约成本，分布式存储系统一般采用数以千计的廉价设备来构建，系统中节点出现故障的概率很高。通常地，在大规模存储系统中，节点出现故障并发生失效已成为一种大概率事件[6]。如图 1.1 所示，Facebook 公司于 2013 年在 ACM 大规模数据库会议（Very-Large Data Bases，VLDB）上发表的一篇文章中指出，在一个具有 3000 个节

点的存储系统中，平均每天有 20 个以上的节点发生故障[7]。为了防止数据因节点失效而丢失，通常将数据按副本（最常用的是三副本）方式存放在不同节点上，从而，当某节点失效时，可以从其他节点获得相同数据。在节点未失效时，副本方式能够支持 I/O 均衡访问和 I/O 并发访问。

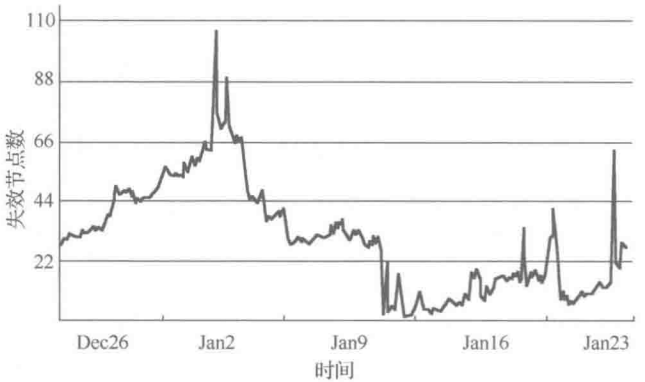


图 1.1　3000 个节点组成的 Facebook 生产集群在一个月期间的失效节点数

副本机制是一种最简单的冗余策略，也称为镜像方法，其基本思想是将数据文件按照固定大小切成分块，每个数据分块在多个位置都保存副本。数据的可靠性与副本数目呈正相关，副本数目越多，数据的可用性越好，可靠性也越高。以 *N*-路副本为例，当不多于 *N*–1 份副本失效时，副本存储系统总能找到一份完好数据将所有失效数据恢复出来。副本技术不涉及编解码运算，读写效率高，实现简单，支持数据并行访问。此外，多个副本可以分担用户访问请求，减少单个节点的 I/O 访问和网络传输，达到负载均衡的效果。

存储效率（storage efficiency）是冗余存储系统的一个重要衡量指标，其具有存储空间利用率的含义，其公式定义为

$$\text{存储效率}=\text{数据空间}/(\text{数据空间}+\text{校验空间}) \tag{1.1}$$

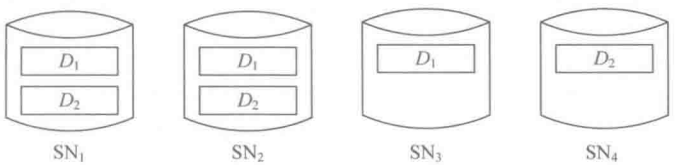
对于三副本存储，其存储效率为 1/3。除了存储效率指标，冗余存储系统还有存储冗余度（storage redundancy）和存储开销（storage overhead）的评价指标。存储冗余度与存储效率互为倒数，其公式定义为

$$\text{存储冗余度}=(\text{数据空间}+\text{校验空间})/\text{数据空间} \tag{1.2}$$

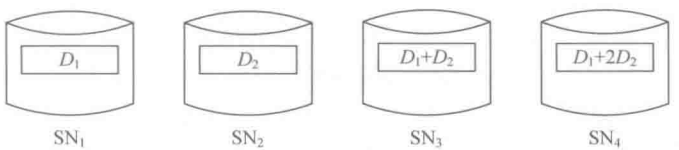
$$\text{存储开销}=\text{校验空间}/\text{数据空间} \tag{1.3}$$

对于三副本存储，其存储冗余度为 3。额外存储开销通常指校验空间和数据空间之间的百分比，此时，三副本的存储开销是 200%。按常理，如果数据按单副本存放，则不存在冗余，即冗余度为 0%，此时，存储冗余度等同于额外存储开销。不同文献对上述指标有不同定义，不管如何定义这些指标，本书统一按式（1.1）～式（1.3）来分析各种冗余存储方案。

对于单个数据中心，三副本存储策略成为业界事实上的容错标准，它能够有效地保证数据可用性，并提高访问并行度。三副本存储策略在存储数据时，将数据复制三份，分别存放在不同的三个节点上，如果其中任意一个或两个节点失效导致数据无法访问，可以从剩余的存活节点读取数据，同时还可以将存活节点上的数据复制到其他节点，以保证系统可靠性。当然，空间利用率很低是副本策略的一大缺点。图 1.2(a)给出了在分布式存储系统中采用副本技术作为容错机制的示意图。该分布式系统中有4个节点，采用三副本的方式保证可以容忍任意两个节点失效。假设有数据分块 D_1 和 D_2 需要存储，每个数据分块需要三个副本，分别存放在不同的节点上，此时系统的空间利用率只有 1/3，也就是增加了 200%的额外存储开销。



(a) 副本方式



(b) 纠删码方式

图 1.2 两种容错方式示意图

1.1.3 纠删码具有高存储效率特性

如上所述，虽然副本技术存在性能优势，但随着存储规模的增大，系统开销（如容量空间成本和运营成本）将显著增加。相对于副本技术，纠删编码技术具有更高存储效率，以图 1.2(b)为例，为了能达到与三副本相同的容错能力（即容忍 2 个节点失效），对数据分块 D_1 和 D_2 进行数据变换，分别是 D_1+D_2 和 D_1+2D_2，此时，总共需要存储 4 个分块，其存储效率为 1/2，即存储效率从三副本的 1/3 提高到 1/2；而存储开销从三副本的 200%降低到 100%。在网络环境下，纠删编码技术的高存储效率这一特性具有更好的表现，它能显著降低网络中的数据流量。因此将纠删编码用于集群存储能节约网络带宽和存储空间，例如，DiskReduce 和 Hadoop-EC 方案将纠删编码应用于 Hadoop 集群存储系统[8,9]。

纠删编码起源于通信传输领域，随着存储可靠性需求的增长，纠删码逐渐应用到存储系统中的数据检错和纠错问题中。在编码参数为（$k+r, k$）的存储系统中，纠删码策略首先将文件数据切分成 k 个数据分块，然后用编码算法对其编码得到 $k+r$ 个编码分块，通过将这些数据分块和冗余块分散到系统中的不同节点，达到容错的目的。如

果该纠删码是具有系统码（systematic codes）特性的里德所罗门编码（Reed Solomon Codes，RS 码）[10]，那么 $k+r$ 个编码分块就包括 k 个数据分块和 r 个冗余分块（也称为校验分块），对于这 $k+r$ 个编码分块，其中任意 r 个分块（包括数据分块和校验分块）失效，可通过相应解码算法计算出 k 个原始数据分块。

当存储系统规模较小时，三副本方式所造成的存储空间开销在可承受范围之内，如 Google、Facebook、Microsoft、Amazon 等公司采用三副本方式来存放其新创建数据（热数据）。当存储系统规模不断增大时，系统的硬件开销和用电能耗将给企业带来沉重的负担。随着数据的爆炸式增长与存储系统规模的日益扩大，副本机制越来越难以满足系统对高空间利用率和高容错度的要求。以数据中心为例，报告[5]显示，Google 在全球的数据中心运行截至 2011 年已超过 90 万台服务器，相应的数据存储开支将达到上亿美元，如果数据中心全部按三副本方式来放置数据，则 2/3 的存储空间用于存放冗余数据（即第二和第三副本），换句话说，大约 30 万台服务器为主副本提供存储空间。如果采用（9, 6）RS 码，那么为原始数据提供存储空间的服务器将提升到 60 万台，只有 1/3 的存储空间用于存放冗余数据（即校验分块）。按式（1.3），三副本的存储开销为 200%，而（9, 6）RS 码的存储开销仅为 50%，二者差距巨大。

简言之，随着数据量的急剧增加，副本存储策略带来的高昂存储开销问题也变得更加突出，为了降低三副本存储策略 200%额外存储开销，各大存储系统逐渐使用纠删码存储策略代替三副本存储策略来存储低访问频度数据。相对于副本技术，纠删码技术在没有过多额外存储空间开销的基础上，通过合理的计算开销和相对低的存储开销来保障存储系统的高可靠性。由于纠删码技术在没有降低系统可靠性的条件下能大幅度减少额外存储空间开销，因此已广泛应用于国内外云存储系统和数据中心。

1.2　纠删码存储的研究现状

纠删码存储是可靠存储的一个重要分支，近年来，与存储相关的国际顶级会议，如 USENIX Conference on File and Storage Technology（FAST）、USENIX Annual Technical Conference （ATC）、IEEE/IFIP Conference on Dependable Systems and Networks（DSN）、ACM Symposium on Operating Systems Principles（SOSP）、USENIX Conference on Operating Systems Design and Implementation（OSDI）等都给予其很高的关注度。

1.2.1　纠删码存储方案的分类

按照存储单元连接方式，纠删码存储可分为基于高速总线方式的磁盘阵列，基于 LAN 方式的集群存储和基于 WAN/Internet 方式的广域网存储系统，如表 1.1 所示。阵列码是一种特殊化的纠删码，它采用高效率的异或运算（XOR），如容单错的 RAID-5 编码[11]和容双错的 RAID-6 编码[12,13]。集群存储中，HDFS-RAID 为 HDFS 增加纠删码支持[7]，PanFS 支持 RAID-5 容错编码[14]，微软 WAS[15]、TAHOE[16]和 Google GFS II[17]