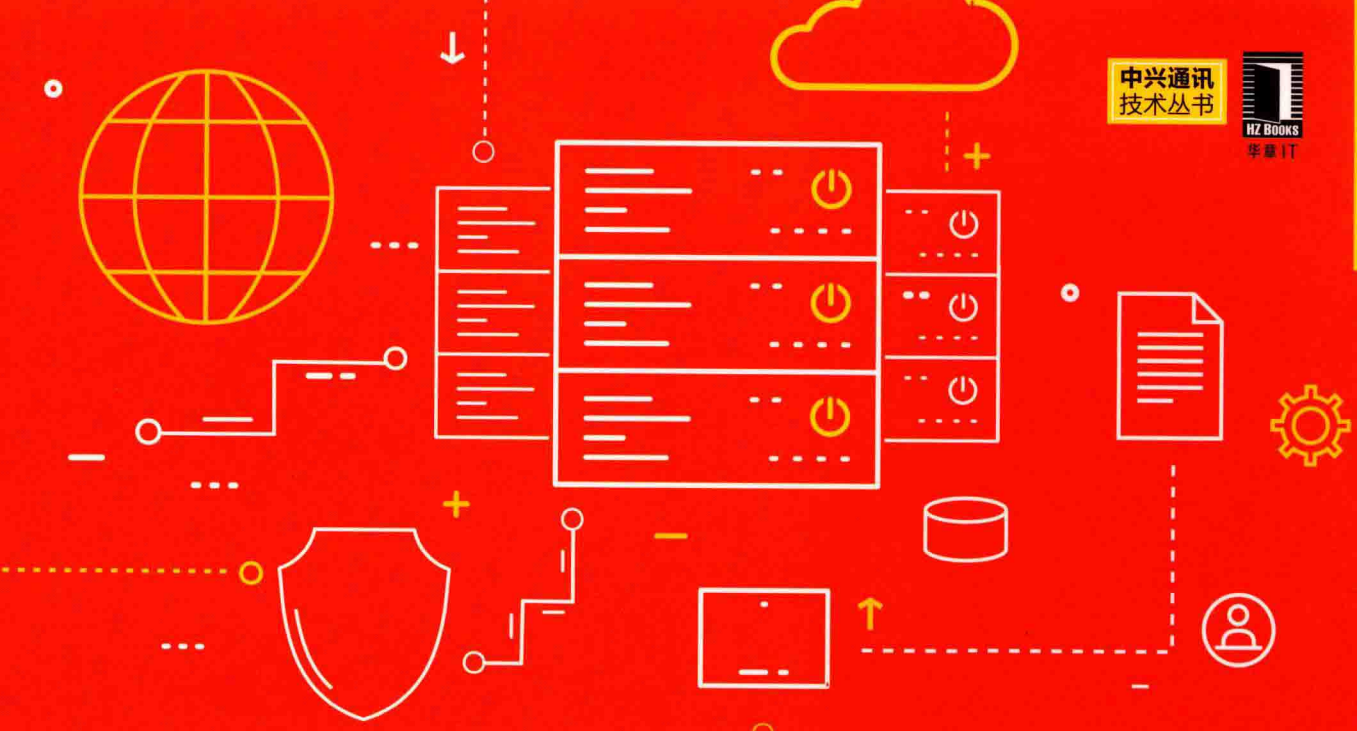
中兴通讯
技术丛书



Ceph

之RADOS设计原理与实现

CEPH RADOS PRINCIPLE AND IMPLEMENTATION

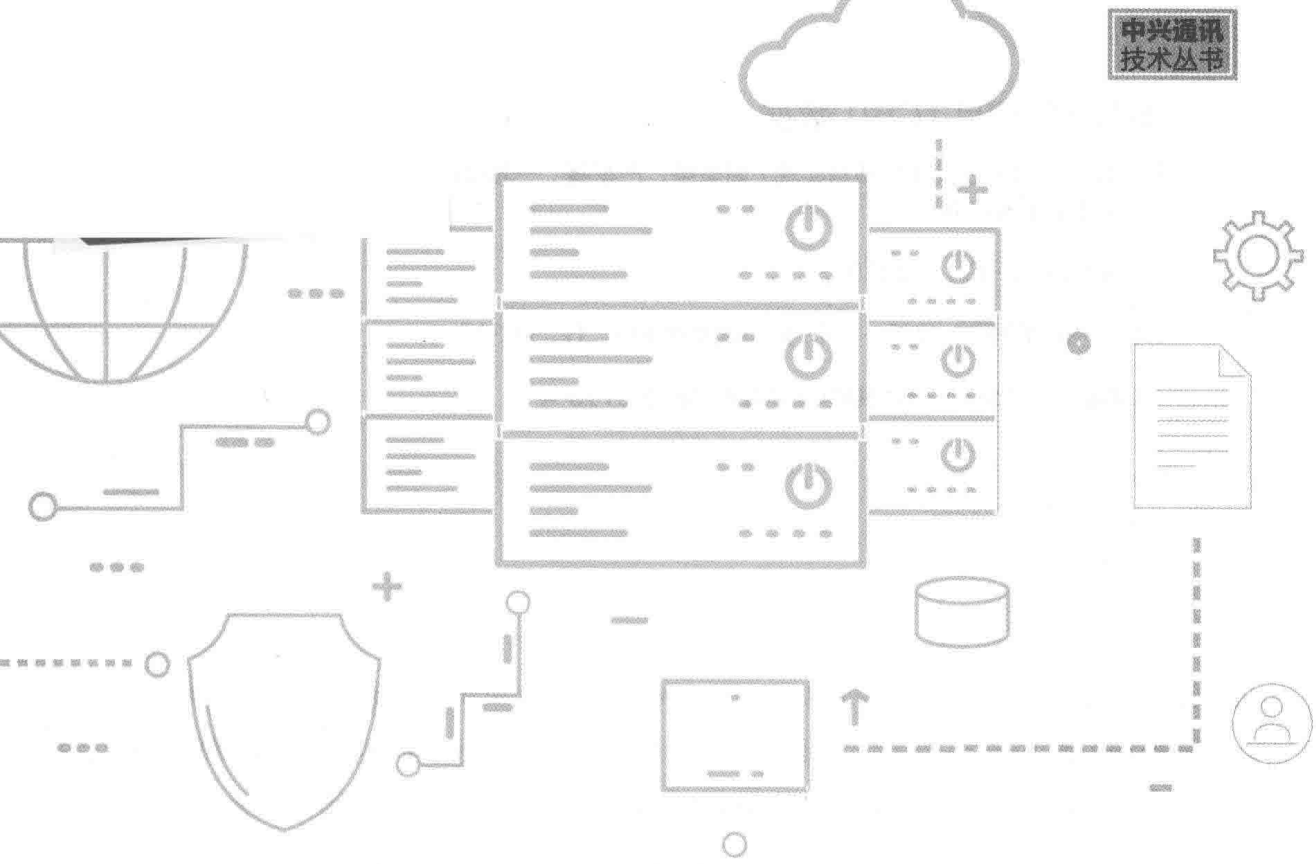
谢型果　严军　◎著

Ceph创始人Sage Weil亲自作序

中兴通讯Clove团队核心成员、Ceph开源社区技术委员会成员与Ceph Member联袂奉献

机械工业出版社
China Machine Press

中兴通讯
技术丛书



Ceph

之RADOS设计原理与实现

CEPH RADOS PRINCIPLE AND IMPLEMENTATION

谢型果　严军　◎著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Ceph 之 RADOS 设计原理与实现 / 谢型果，严军著 . —北京：机械工业出版社，2019.1
（中兴通讯技术丛书）

ISBN 978-7-111-61389-3

I. C… II. ①谢… ②严… III. 分布式文件系统 IV. TP316

中国版本图书馆 CIP 数据核字（2018）第 263710 号

Ceph 之 RADOS 设计原理与实现

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：陈佳媛 责任校对：殷 虹
印 刷：中国电影出版社印刷厂 版 次：2019 年 1 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16 印 张：18.5
书 号：ISBN 978-7-111-61389-3 定 价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzit@hzbook.com

华章IT

HZBOOKS | Information Technology

It has been my pleasure to see the growth in popularity of Ceph in China over the last several years. More than any other region, I see organizations recognizing the benefits of an open source, distributed, and scalable storage platform for file, block, and object workloads. Open source is a huge opportunity for organizations to work cooperatively to improve software and to benefit from each others' hard work.

The Chinese Ceph community has become an impressive force in the larger Ceph ecosystem. In March of 2018 the first Cephalocon conference took place in Beijing. The event spanned two days, included more than 60 speakers, and drew more than 1 000 attendees. I was incredibly impressed with the amount of technical content that was presented by the local community, and with the scope of Ceph deployments in China. I look forward to returning soon for the next Cephalocon or Ceph Day event!

My primary goal is to grow and build the Ceph developer community in China. As more organizations rely on Ceph for their storage infrastructure, the pool of serious Ceph users grows, and every user presents a new opportunity to become involved in Ceph development. We face a number of challenges that make it hard for Chinese developers to participate in the upstream Ceph development community, including language, reliable and unrestricted internet access to common collaboration tools (including chat, video conferencing, and video archives), and time zones. However, I think these challenges can be overcome by building a community nexus that

is centered in China, using what languages and tools are most convenient and natural, and by facilitating communication and collaboration with the broader community through experienced developers like Xie Xingguo. Books like this one that help new developers become familiar with Ceph are a critical part in this effort.

For the Ceph project I see a few key focus area for the next few years: (1) usability, (2) performance, (3) integration with emerging container platforms (like Kubernetes), and (4) hybrid and multi-cloud capabilities.

Ceph has developed a reputation for being hard, and that has slowed adoption. This is a problem we need to fix—not only to make Ceph accessible to a broader set of users, but also to make the system manageable at large scales by small teams of operators. As storage demands grow, the scale of storage infrastructure will grow too, and making Ceph automate as much of its day to day operations as possible will become increasingly important.

The storage hardware landscape is also changing. Ceph is pure software, which means it can run of a broad range of systems and storage devices. However, as the industry continues to shift to solid state storage technologies like NAND flash and persistent memory technologies, the software will need to evolve and adapt to capture the performance of that hardware. A critical strategic effort is now underway to reimplement key parts of the Ceph data path using new software technologies, including SeaStar, SPDK, and DPDK. The success of this effort will depend on the participation and efforts of many community members.

The Kubernetes container platform is increasingly looking like it will dominate the next wave of IT infrastructure, and with any scale-out infrastructure platform, scale-out software-defined storage technologies will be needed. The Ceph community is working hard to ensure that Ceph is the easiest and obvious choice for storage in this growing ecosystem through projects like Rook, but with any growing and evolving technology, the needs are changing all of the time. Sustained engagement with this community—and other emerging infrastructure projects—is needed to make sure we can meet the needs of users.

Finally, modern enterprises are increasingly deploying infrastructure across a range

of different data centers, clouds, geographic regions, and regulatory regimes. Simply storing data reliably within a single cluster is no longer sufficient to solve businesses' storage problems. Federation capabilities, disaster recovery, and data management services that enable portability of applications—and their data—across clouds and data centers is necessary to provide users with the freedom from being locked-in to specific cloud providers or enterprise platforms. The next phase of evolution for Ceph will be in providing the underlying storage features that enable replication and migration of data sets across clusters and clouds for file, block, and especially object storage. Object storage in particular will be the preferred interface for the next generation of " cloud native " applications, and presents a new opportunity for Ceph to provide higher-level data services for managing the placement, replication, tiering, and migration of data across clouds in an automated, policy-driven way.

More than ever before I am excited about what is coming next for Ceph and the future of storage. I hope that you can be part of that journey!

Sage Weil，Ceph 创始人

序 2 Preface

随着分布式源代码托管网站 GitHub 风靡全球，开源正以燎原之势席卷整个软件世界，成为软件工程的新信条。

开源（软件活动）诞生了众多世界级的明星项目，其中最负盛名的莫过于 Linux，它不仅在构建大型数据中心所必需的服务器操作系统领域占据统治地位，而且也是举世闻名的 Android 智能手机操作系统的核心。可以说 Linux 的出现在很大程度上加速了信息时代，特别是大数据时代的到来。另一个例子则是近年来闪电崛起、已成为云计算事实上标准的 OpenStack。云计算的普及，在不断降低企业生产成本的同时也成倍提升着企业的生产效率。而 OpenStack 不但在私有云领域独领风骚，在公有云领域，全球有超过 60 个公有云也采用或者借鉴了它的技术架构。随着开源力量的进一步壮大，可以预见，这样的例子将会越来越多。

除了明星开源软件对于整个行业乃至整个世界产生的羊群效应，开源本身也代表了一种先进的软件开发与管理理念。

首先，开源代表开放、透明、高效、兼容并包与拒绝重复。由于借助的是全世界开发者（也包含最终用户）的力量，开源产品可以获得最广泛的需求收集、技术讨论与用户体验及反馈渠道，也得以最大程度地避免无效或者重复劳动。因此开源产品更新换代速度快、通用性好并且质量可靠，同时由于开放源码，可以保证用户隐私不被窃取，这在“数据即价值”的大数据时代显得尤为可贵。

其次，开源社区人才辈出、大师云集。与大师同行，接受大师熏陶，与开源社区共成长，是快速提升自身的不二法门。

最后，开源社区的运作理念，例如提倡高度自动化、自管理、自组织等，其实是一种更高层次的敏捷。以开源社区的方式进行产品研发，无论对于产品本身，还是对于参与其中的每个人，无疑都会大有裨益。

中兴通讯深刻认同并重视开源的力量。多年来，我们持续投入并深耕于包括OpenStack、OpenNFV、Kubernetes、Ceph等在内的多个与云计算相关的全球性主流开源组织，极大地影响着，甚至引导着相关行业标准制定，成为推动行业变革、加速行业进化的关键性力量。作为中兴通讯开源领域的旗帜人物，本书作者之一谢型果是Ceph开源社区的技术委员会成员，他所带领的团队一直奋斗在社区最前沿，连续多个版本代码贡献数量位列前三。同时，中兴通讯基于Ceph的分布式存储产品商用规模也已经超过10 PB。以Ceph为代表，中兴通讯基于开源打造的系列产品经过理论与实践双重检验，真正做到了源于开源、高于开源，同时不忘回馈开源。

开源改变世界，我很荣幸能够见证这一历史时刻的到来并参与其中。

杨日　中兴通讯无线研究院副院长兼大数据研发中心主任

序　3 Preface

随着 ICT 产业不断融合与发展，传统电信运营商开始逐步采用云计算、虚拟化、软件定义网络等 IT 领域新技术优化通信网络架构，通过网络功能虚拟化（NFV）等方式对软、硬件进行解耦，以提高资源利用率，实现网元快速部署与升级、降低维护和运营成本等目标，也为即将到来的 5G 移动通信提供更具弹性的基础设施平台。

作为传统 CT 领域的重要设备厂商，中兴通讯在当今开放、共享、数字经济转型的时代大背景下，始终重视并持续关注新技术的发展，积极推动 ICT 产业技术融合，致力于为客户提供更加优质的数字化与信息化服务。在寻求自身技术转型的过程中，中兴通讯十分注重云计算相关基础设施的建设。Ceph 作为当今最先进的分布式统一存储平台之一，已成为私有云 /NFVI 事实上的标准——OpenStack 的默认存储后端，也是中兴通讯重点关注和投入的方向。

《 Ceph 之 RADOS 设计原理与实现》是中兴通讯第二本原创性 Ceph 著作。中兴通讯 Ceph 团队从传统存储研发转型，近年来聚焦于以 Ceph 为代表的分布式存储领域，在与开源社区保持紧密协作的同时，高效支撑了公司国内外多个电信云商用局点的存储解决方案，具备较高的理论水平和丰富的实践经验。两位作者同为 Ceph 社区的核心贡献者（Ceph Member），长期工作在社区一线，无论是贡献数量还是质量都在国内首屈一指。这次他们能够付出大量时间和精力将多年来的研究心得与实践经验编撰成册，非常难能可贵。

本书理论结合实际，深入浅出地介绍了 Ceph 系统的核心组件——RADOS 的设计理念与实现方式，相信无论是从事 Ceph 研发还是运维工作的读者，都将从中获益匪浅。

陈新宇　中兴通讯电信云及核心网产品副总经理（主管研发）

Preface 前　言

2018 年 3 月，全球 Cephers 的盛会——Cephalocon APAC 2018 在北京举行。来自 RedHat、SUSE、Intel、中兴通讯、中国移动等 Ceph 生态联盟成员的 1000 多位 Ceph 开发者、使用者和爱好者共聚一堂，探讨 Ceph 的现状与未来，彰显了 Ceph 开源社区的蓬勃生机。

时光荏苒，自 Ceph 由 Sage A. Weil 在博士论文提出以来，十多年间，已经从一个默默无闻的学生作品成长为分布式存储领域最具活力与领导力的开源项目。据 Ceph 官方不完全统计，在世界范围内，目前已有超过 100 家公司（机构）研究与使用 Ceph，其中不乏欧洲原子能研究组织（CERN）这样知名的全球性科研机构和 Yahoo、阿里巴巴等著名的互联网公司。可见，作为分布式软件定义存储的标杆，Ceph 领先的架构和设计理念已经深入人心。

Ceph 的魅力源于其架构的前瞻性、可塑性和长期演进能力。事实上，在设计之初，Ceph 被定位成一个纯粹的分布式文件系统，主要用于解决大型超级计算机之间如何通过联网的方式提供可扩展的文件存储服务。随着云计算、大数据和人工智能逐渐成为信息时代的主旋律，Ceph 正不断拓展自身的触角，从取代 Swift 成为 OpenStack 首选存储后端进入公众视野，到完美适配以 Amazon S3 为代表的公有云接口，再到征战下一个没有硝烟的虚拟化（技术）高地——容器。时至今日，Ceph 已然成为一个兼容块、文件、对象等各类经典 / 新兴存储协议的超级统一存储平台。随着 Ceph 的加速进化，可以预见，我们将会看到越来越多的基于 Ceph 构建的自定义存储应用。

为什么写这本书

开源软件诞生的土壤决定了大部分开源软件从来就不是面向普通大众的，典型的如 Linux，其无可视化界面的命令行操作方式和海量命令足以让 90% 的用户望而却步[⊖]。Ceph 作为一个出身于学院的开源作品也存在类似的缺点[⊜]。此外，随着自身的不断演进和完善，Ceph 已经从最初的分布式文件系统逐渐成长为一个全能的分布式统一存储平台，因此其复杂程度远远超过功能相对单一的传统存储系统[⊜]。更糟的是，虽然社区有建议的编码规范，但是为了不挫伤贡献者的积极性，这些规范并未作为强制要求，因此随着贡献者数量的快速增长，Ceph 代码本身也不可避免地趋于异构化。上述种种因素使得无论是使用还是开发 Ceph 都难度巨大，再加上语言和文化背景的差异，足以造成大量国内 Ceph 初级玩家难以逾越的鸿沟。

距我们创作《Ceph 设计原理与实现》[⊛]一书已经过去了两年。一方面，Ceph 代码发生了巨大变化；另一方面，我们对 Ceph 的认知也有了较大提升。因此，我们两位负责研究 RADOS 组件的同事基于前作中的相关章节重新创作了本书。

与前作相比，本书更加专注于 RADOS 这个基础组件，而剥离了 RBD、RGW、CephFS 等具体存储应用和案例实战部分。这主要是基于以下考虑：

首先，RBD、RGW 和 CephFS 与其承载的具体业务耦合度较高，例如 RBD 后续的重点工作是兼容 iSCSI/FC 传统块存储接口，而要彻底掌握 RGW 则必然要对以 S3、Swift 为代表的新兴对象存储协议簇有比较透彻的了解等，限于篇幅，很难单纯从 Ceph 的角度对这些组件做出比较完整和透彻的解读。

其次，由于时间仓促，加之不少章节均由不同的作者独立创作，因此前作中章节之间难免重复或者脱节，而本书则更加注重章节之间衔接与编排的合理性。此外，由于作者数量大幅减少，本书风格更加统一，相对而言读者可以获得更好的阅读体验。

⊖ 近年来，随着 Linux 各种自带桌面应用版本（例如 Ubuntu）的流行，这种情况有所好转。

⊜ 在 2016 年 RedHat Summit 上，有人问 Sage："Ceph 当前最大的缺点是什么？" Sage 回答："Hard to use." 此外，在 2018 年 Cephalocon APAC 2018 峰会上，Sage 在主题演讲中再次将提升 Ceph 的易用性作为后续发展的战略目标之一（位列性能之后）。

⊜ 作为对比，被誉为最后一个本地文件系统的 ZFS（采用 C 语言开发）和 Ceph（采用 C++ 开发）的代码量之比约为 1 ： 10。

⊛ 由机械工业出版社出版，书号为 978-7-111-57842-0。

再次，藉本次重新创作，我们进一步削弱了前作中相关章节与代码之间的耦合性，更加侧重于阐述设计理念。由于 Ceph 社区十分活跃，贡献者数量众多，每个版本代码都会发生翻天覆地的变化，因此，理解设计原理，以不变应万变，无疑比掌握某个特定版本的代码更为重要。

最后，需要再次强调的是，虽然本书部分章节源自《Ceph 设计原理与实现》一书，但是基本上都进行了重新创作。重复录入这些章节不是简单的查漏补缺，而是进一步提炼与升华，它们是本书不可或缺的组成部分。事实上，与新增内容相比，重新创作这些章节花费了我们更多的时间与精力。

本书的读者对象

本书适合于对 Ceph 有一定了解，想更进一步参与到 Ceph 开源项目中来，并致力于后续为 Ceph，特别是 RADOS 组件添砖加瓦的开发者或者高级开发者阅读。

此外，高级运维人员通过阅读本书也能够了解和掌握 Ceph 的核心设计理念及高级应用技巧，从而在日常运维工作中更加得心应手。

与《Ceph 设计原理与实现》力求如实反映源码的实现细节不同，本书是 Ceph（特别是 RADOS 组件）设计思想与基本理念的高度浓缩。有条件的读者可以将两本书对照阅读，相信可以有更大收获。

本书的主要内容

本书主要介绍 Ceph 的核心——RADOS。具体编排如下：

第 1 章　一生万物——RADOS 导论

Ceph 是集传统块、文件存储以及新兴对象存储于一身的超级分布式统一存储平台。

Ceph 在架构上采用存储应用与存储服务完全分离的模式，并基于 RADOS 对外提供高性能和可轻松扩展的存储服务。理论上，基于 RADOS 及其派生的 librados 标准库可以开发任意类型的存储应用，典型的如 Ceph 当前的三大核心应用：RBD、RGW 和 CephFS。

作为全书的开始，本章旨在为读者建立一个 RADOS 的初步印象，主要介绍包括 OSD、Monitor、存储池、PG、对象等在内的一众基本概念。

第 2 章　计算寻址之美与数据平衡之殇——CRUSH

CRUSH 是 Ceph 两大核心设计之一。CRUSH 良好的设计理念使其具有计算寻址、高并发和动态数据均衡、可定制的副本策略等基本特性，进而能够非常方便地实现诸如去中心化、有效抵御物理结构变化并保证性能随集群规模呈线性扩展、高可靠等高级特性，因而非常适合 Ceph 这类对可扩展性、性能和可靠性都有严苛要求的大型分布式存储系统。

CRUSH 最大的痛点在于，在实际应用中，很容易出现由于 CRUSH 的先天缺陷导致 PG 分布不均，进而导致集群出现 OSD 之间数据分布失衡、集群整体空间利用率不高的问题，为此社区引入了包括 reweight、weight-set、upmap、balancer 在内的一系列手段加以改进。

第 3 章　集群的大脑——Monitor

Monitor 是基于 Paxos 兼职议会算法构建的、具有分布式强一致性的小型集群，主要负责维护和传播集群表的权威副本。Monitor 采用负荷分担的方式工作，因此，任何时刻、任意类型的客户端或者 OSD 都可以通过和集群中任意一个 Monitor 进行交互，以索取或者请求更新集群表。基于 Paxos 的分布式一致性算法可以保证所有 Monitor 的行为自始至终都是正确和自洽的。

第 4 章　存储的基石——OSD

对象存储起源于传统的 NAS（例如 NFS）和 SAN 存储，其基本思想是赋予底层物理存储设备（例如磁盘）一些 CPU、内存资源等，使之成为一个抽象的对象存储设备（即 OSD），能够独立完成一些低级别的文件系统操作（例如空间分配、磁盘 I/O 调度等），以实现客户端 I/O 操作（例如读、写）与系统调用（例如打开文件、关闭文件）之间的解耦。

与传统对象存储仅仅赋予 OSD 一些初级的“智能”不同，Ceph 开创性地认为，这种“智能”可以被更进一步地用于执行故障恢复与数据自动平衡、提供完备的高性能本

地对象存储服务等复杂任务上，从而使得基于 OSD 构建高可靠、高可扩展和高并发性能的大型分布式对象存储系统成为可能。

第 5 章　高性能本地对象存储引擎——BlueStore

BlueStore 是默认的新一代高性能本地对象存储引擎。BlueStore 在设计中充分考虑了对下一代全 SSD 以及全 NVMe SSD 闪存阵列的适配，增加了数据自校验、数据压缩等热点增值功能，面向 PG 提供高效、无差异[⊖]和符合事务语义的本地对象存储服务。

第 6 章　移动的对象载体——PG

面向分布式的设计使得 Ceph 可以轻易管理拥有成百上千个节点、PB 级以上存储容量的大规模集群。

通常情况下，对象大小是固定的。考虑到 Ceph 随机分布数据（对象）的特性，为了最大程度地实现负载均衡，不会将对象粒度设计得很大，因此即便一个普通规模的 Ceph 集群，也可以存储数以百万计的对象，这使得直接以对象为粒度进行资源和任务管理的代价过于昂贵。

简言之，PG 是一些对象的集合。引入 PG 的优点在于：首先，集群中 PG 数量经过人工规划因而严格可控（反之，集群中对象的数量则时刻处于变化之中），这使得基于 PG 精确控制单个 OSD 乃至整个节点的资源消耗成为可能；其次，由于集群中 PG 数量远远小于对象数量，并且 PG 的数量和生命周期都相对稳定，因此以 PG 为单位进行数据同步或者迁移等，相较于直接以对象为单位而言，难度更小。

PG 最引人注目之处在于其可以在 OSD 之间（根据 CRUSH 的实时计算结果）自由迁移，这是 Ceph 赖以实现自动数据恢复、自动数据平衡等高级特性的基础。

第 7 章　在线数据恢复——Recovery 与 Backfill

在线数据恢复是存储系统的重要研究课题之一。

与离线恢复不同，在线数据恢复的难点在于数据本身一直处于变化之中，同时在生产环境中一般都有兼顾数据可靠性和系统平稳运行的要求，因此如何合理地处理各种业

⊖ 指与 FileStore 等传统本地对象存储引擎相比。

务之间的冲突，恰当地分配各种业务之间的资源（例如 CPU、内存、磁盘和网络带宽等），在尽可能提升数据恢复速度的同时，降低甚至完全避免数据恢复对正常业务造成干扰则显得至关重要。

按照能否依据日志进行恢复，Ceph 将在线数据恢复细分为 Recovery 和 Backfill 两种方式。通常意义上，两者分别用于应对临时故障和永久故障，当然后者也常用于解决由于集群拓扑结构变化导致的数据和负载不均衡问题。

第 8 章　数据正确性与一致性的守护者——Scrub

Scrub 是一种重要的辅助机制，用于守护集群数据的正确性与一致性。

实现上，Scrub 主要依赖对象的有序性与信息摘要技术，前者使其可以不重复（从而高效）地遍历集群中的所有对象，后者则提供了一种快速检测数据正确性和一致性的手段。

与数据恢复类似，Scrub 也分在线和离线两种方式。同样，由于数据本身一直处于变化之中，为了捕获数据错误和一致性问题，要求 Scrub 周期性地执行，同时为了能够完整地完成一次深度扫描，则要求 Scrub 基于合适的粒度、以合理的规则执行。

第 9 章　基于 dmClock 的分布式流控策略

dmClock 是一种基于时间标签的分布式 I/O 调度算法。Ceph 采用 dmClock 主要希望解决客户端业务与集群内部操作的 I/O 资源合理分配问题，但由于种种原因，这一部分研究进展比较缓慢。通过深入研究 dmClock 算法，我们在社区的基础上对其进行了改进和应用实践，增加了块设备卷粒度 QoS 与 OSD 粒度的数据恢复流量自适应控制的支持，并基于此进一步优化了整个集群的流控策略。

第 10 章　纠删码原理与实践

Ceph 传统的 3 副本数据备份方式能够在取得高可靠性的前提下最小化客户端请求的响应时延，因而特别适合对可靠性和性能都有一定要求的存储应用。这种目前使用最广泛的备份方式的缺点在于会大量占用额外的存储空间，因而导致集群的实际空间利用率不高。与之相反，纠删码以条带为单位，通过数学变换，将采用任意 $k+m$ 备份策略所消

耗的额外存储空间都成功控制在 1 倍以内，而代价是计算资源消耗变大和客户端请求响应时延变长，因而适合对时延不敏感的“冷数据”（例如备份数据）应用。

勘误与支持

《Ceph 设计原理与实现》一书出版之后，我们收到了不少读者朋友的来信，指出了书中的错误和疏漏，在此对这些热心的读者朋友们一并表示感谢。

本书在前作的基础上，对相关章节进一步做了大幅修订，同时增加了大量新的章节。由于写作和认知水平有限，问题仍然在所难免，欢迎新老读者们通过以下电子邮箱对本书进行指正：

xie.xingguo@zte.com.cn

yan.jun8@zte.com.cn

致谢

Ceph 官方社区[⊖]的源代码[⊜]是创作本书的原始素材，因此我们首先感谢 Ceph 官方社区和创始人 Sage A. Weil 先生。

其次，我们要感谢所在部门的主管领导谭芳，感谢他在日常工作中给予我们的关照，以及对于出版《Ceph 设计原理与实现》与本书不遗余力的支持。

再次，感谢 Clove 团队，本书很大程度上是整个团队的智慧结晶。此外，我们要特别感谢宋维斌、韦巧苗、罗慕尧、朱尚忠、朱凯波等几位同事，他（她）们通读了本书的初稿，提出了大量宝贵的修改意见，极大地增强了本书的专业性与可读性。

最后，感谢 IT 学院的闫林老师对于出版《Ceph 设计原理与实现》及本书给予的鼓励和巨大帮助。

谢型果　严军

2018 年 8 月

⊖ https://ceph.com/

⊜ https://github.com/ceph/ceph