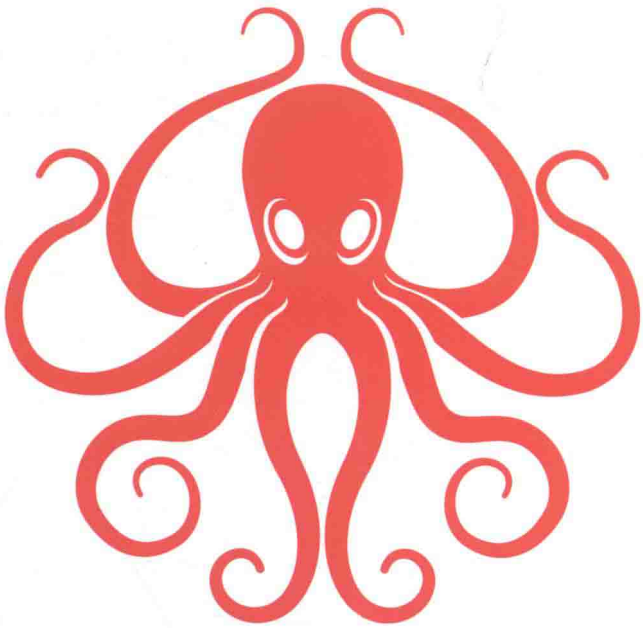
从源码角度全面解析Ceph的整体框架和各个模块的实现原理

包含作者多年开发经验，掌握分布式存储技术必备参考

HZ BOOKS
华章科技

大数据技术丛书



The Source Code Analysis of Ceph

Ceph源码分析

常涛◎编著

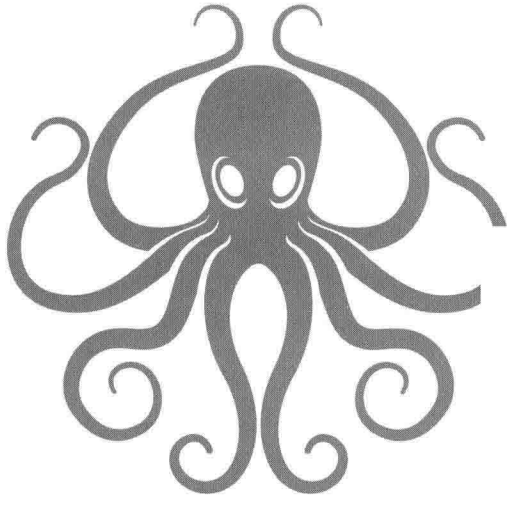
机械工业出版社
China Machine Press

大数据技术丛书

The Source Code Analysis of Ceph

Ceph源码分析

常涛◎编著



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Ceph 源码分析 / 常涛编著 . —北京：机械工业出版社，2016.10
（大数据技术丛书）

ISBN 978-7-111-55207-9

I. C… II. 常… III. 分布式文件系统 IV. TP316

中国版本图书馆 CIP 数据核字（2016）第 257284 号

Ceph 源码分析

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：吴 怡 责任校对：殷 虹
印 刷：北京市荣盛彩色印刷有限公司 版 次：2016 年 11 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16 印 张：16.75
书 号：ISBN 978-7-111-55207-9 定 价：59.00 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

序　言

自从 2013 年加入 Ceph 社区以来，我一直想写一本分析 Ceph 源码的书，但是两年多来提交了数万行的代码后，我渐渐放下了这个事情。Ceph 每个月、每周都会发生巨大变化，我总是想让 Ceph 源码爱好者看到最新最棒的设计和实现，社区一线的模块维护和每周数十个代码提交集的阅读，让我很难有时间回顾和把握其他 Ceph 爱好者的疑问和需求点。

今天看到这本书让我非常意外，作者常涛把整个 Ceph 源码树肢解得恰到好处，如庖丁解牛般将 Ceph 的核心思想和实现展露出来。虽然目前 Ceph 分分钟都有新的变化，但无论是新的模块设计，还是重构已有逻辑，都是已有思想的翻新和延续，这些才是众多 Ceph 开发者能十年如一日改进的秘密！

我跟作者常涛虽然只有一面之缘，但是在开源社区中的交流已经足够成为彼此的相知。他对于分布式存储的设计和实现都有独到见解，其代码阅读和理解灵感更是超群。我在年前看到他一些对 Ceph 核心模块的创新性理解，相信这些都通过这本书展现出来了。

这本书是目前我所看到的从代码角度解读 Ceph 的最好作品，即使在全球范围内，都没有类似的书籍能够与之媲美。相信每个 Ceph 爱好者都能从这本书中找到自己心中某些疑问的解答途径。

作为 Ceph 社区的主要开发者，我也想在这里强调 Ceph 的魅力，希望每个读者都能充分感受到 Ceph 社区生机勃勃的态势。Ceph 是开源世界中存储领域的一个里程碑！在过去很难想像，从 IT 巨无霸们组成的巨大存储壁垒中能够诞生一个真正被大量用户使用并投入生产环境的开源存储项目，而 Ceph 这个开源存储项目已经成为全球众多海量存储项目的主要选择。

众所周知，在过去十年里，IT 技术领域中巨大的创新项目很多来自于开源世界，从垄断大数据的 Hadoop、Spark，到风靡全球的 Docker，都证明了开源力量推动了新技术的产生与发展。而再往以前看十年，从 Unix 到 Linux，从 Oracle 到 MySQL/PostgreSQL，从 VMWare 到 KVM，开源世界从传统商业技术继承并给用户带来更多的选择。处于开源社区一线的我欣喜地看到，在 IT 基础设施领域，越来越多的创业公司从创立之初就以开源为基石，而越来越多的商业技术公司也受益于开源，大量的复杂商业软件基于开源分布式数据库、缓存存储、中间件构建。相信开源的 Ceph 也将成为 IT 创新的驱动力。正如 Sage Weil 在 2016 Ceph Next 会议上所说，Ceph 将成为存储里的 Linux！

王豪迈，XSKY 公司 CTO

2016 年 9 月 8 日

前　言

随着云计算技术的兴起和普及，云计算基石：分布式共享存储系统受到业界的重视。Ceph 以其稳定、高可用、可扩展的特性，乘着开源云计算管理系统 OpenStack 的东风，迅速成为最热门的开源分布式存储系统。

Ceph 作为一个开源的分布式存储系统，人人都可以免费获得其源代码，并能够安装部署，但是并不等于人人都能用起来，人人都能用好。用好一个开源分布式存储系统，首先要对其架构、功能原理等方面有比较好的了解，其次要有修复漏洞的能力。这些都是在采用开源分布式存储系统时所面临的挑战。

要用好 Ceph，就必须深入了解和掌握 Ceph 源代码。Ceph 源代码的实现被公认为比较复杂，阅读难度较大。阅读 Ceph 源代码，不但需要对 C++ 语言以及 boost 库和 STL 库非常熟悉，还需要有分布式存储系统相关的基础知识以及对实现原理的深刻理解，最后还需要对 Ceph 框架和设计原理以及具体的实现细节有很好的把握。所以 Ceph 源代码的阅读是相当有挑战性的。

本着对 Ceph 源代码的浓厚兴趣以及实践工作的需要，需要对 Ceph 在源代码层级有比较深入的了解。当时笔者尽可能地搜索有关 Ceph 源代码的介绍，发现这方面的资料比较少，笔者只能自己对着 Ceph 源代码开始了比较艰辛的阅读之旅。在这个过程中，每一个小的进步都来之不易，理解一些实现细节，都需要对源代码进行反复地推敲和琢磨。自己在阅读的过程中，特别希望有人能够帮助理清整体代码的思路，能够解答一下关键的实现细节。本书就是秉承这样一个简单的目标，希望指引和帮助广大 Ceph 爱好者更好地理解和掌握 Ceph 源代码。

本书面向热爱 Ceph 的开发者，想深入了解 Ceph 原理的高级运维人员，想基于 Ceph 做优化和定制的开发人员，以及想对社区提交代码的研究人员。官网上有比较详细的介

绍 Ceph 安装部署以及操作相关的知识，希望阅读本书的人能够自己动手实践，对 Ceph 进一步了解。本书基于目前最新的 Ceph 10.2.1 版本进行分析。

本书着重介绍 Ceph 的整体框架和各个实现模块的实现原理，对核心源代码进行分析，包括一些关键的实现细节。存储系统的实现都是围绕数据以及对数据的操作来展开，只要理解核心的数据结构，以及数据结构的相关操作就可以大致了解核心的实现和功能。本书的写作思路是先介绍框架和原理，其次介绍相关的数据结构，最后基于数据结构，介绍相关的操作实现流程。

最后感谢一起工作过的同事们，同他们在 Ceph 技术上进行交流沟通并加以验证实践，使我受益匪浅。感谢机械工业出版社的编辑吴怡对本书出版所做的努力，以及不断提出的宝贵意见。感谢我的妻子孙盛南女士在我写作期间默默的付出，对本书的写作提供了坚强的后盾。

由于 Ceph 源代码比较多，也比较复杂，写作的时间比较紧，加上个人的水平有限，错误和疏漏在所难免，恳请读者批评指正。有任何的意见和建议都可发送到我的邮箱 changtao381@163.com，欢迎读者与我交流 Ceph 相关的任何问题。

常涛

2016 年 6 月于北京

目　录

第 1 章 Chapter 1

Ceph 整体架构

本章从比较高的层次对 Ceph 的发展历史、Ceph 的设计目标、整体架构进行简要介绍。其次介绍 Ceph 的三种对外接口：块存储、对象存储、文件存储。还介绍 Ceph 的存储基石 RADOS 系统的一些基本概念、各个模块组成和功能。最后介绍了对象的寻址过程和数据读写的原理，以及 RADOS 实现的数据服务等。

1.1 Ceph 的发展历程

Ceph 项目起源于其创始人 Sage Weil 在加州大学 Santa Cruz 分校攻读博士期间的研究课题。项目的起始时间为 2004 年，在 2006 年基于开源协议开源了 Ceph 的源代码。Sage Weil 也相应成立了 Inktank 公司专注于 Ceph 的研发。在 2014 年 5 月，该公司被 Red Hat 收购。Ceph 项目的发展历程如图 1-1 所示。

2012 年，Ceph 发布了第一个稳定版本。2014 年 10 月，Ceph 开发团队发布了 Ceph 的第七个稳定版本 Giant。到目前为止，社区平均每三个月发布一个稳定版本，目前的最新版本为 10.2.1。