Quick answers to common problems

Ceph Cookbook

Over 100 effective recipes to help you design, implement, and manage the software-defined and massively scalable Ceph storage system

Foreword by Dr. Wolfgang Schulze
Director of Global Storage Consulting, Red Hat

Karan Singh

[PACKT] PUBLISHING open source community experience distilled

Table of Contents

Ceph Cookbook

超过 100 个有效的方法来帮助你设计，实施和管理软件定义存储，大规模的扩展 Ceph 存储系统

Karan Singh

Ceph Cookbook

First published: February 2016
Production reference: 1250216
Published by Packt Publishing Ltd.Livery Place
35 Livery Street
Birmingham B3 2PB, UK

ISBN 978-1-78439-350-2

参与人员

角色	人员
作者	Karan Singh
评审	Christian Eichelmann Haruka Iwao
策划编辑	Amarabha Banerjee
选稿编辑	Meeta Rajani
内容发掘编辑	Kajal Thapar
技术编辑	Menza Mathew
审稿	Angad Singh
项目协调	Shweta H Birwatkar
校对	Safis Editing
目录	Rekha Nair
产品协调	Melwyn Dsa
封面	Melwyn Dsa

前言

一年前，Karan 出版了他的第一本书，Learning Ceph，Packt 发行的，取得了非常大的成功。他解决了很多用户的需求：介绍 ceph 和 ceph 架构的概览，并且能够容易理解的书籍。

当一个开源的项目像 Ceph 这样火热的时候，功能的迭代就会非常的迅速。除了以 sage 带头的红帽开发团队以外，行业巨头中，Intel、Sandisk、Fujitsu、Suse以及无数其它的小公司和个人都做出了自己的贡献。其结果就是，这个项目无论是性能上还是稳定性上都变得更加成熟；这些公司在企业部署中起到了很关键的作用，很多功能在上一本书中还处于开发起步阶段，到现在，已经成为 Ceph 的一部分了；纠删码，固态硬盘的优化，VSM 管理平台，还有很多其它的东西，所有的这些将在这本书里面详细的讲述到。

有一天，我读到一篇博客，讲的是，Ceph 对存储行业的影响，就像linux对操作系统的影响。虽然现在这样说还为时尚早，但是 Ceph 通过行业的使用情况证明了这一点，PB级别的部署越来越多。大型的集群部署，例如 CERN 和 Yahoo，定期与社区分享他们的经验。

丰富的功能和极大的灵活性，以适应广泛的案例，有时让人难以接近这种新技术，他会让新手不知道从哪里开始学习。不是每个人都可以访问到数千台的服务器和硬盘进行实验，并且创建数据中心。Karan 的新书 Ceph Cookbook 将会通过实践，以及之前的经验来教你怎么面对的这些挑战。

作为一个长期的ceph爱好者，我曾经与 Karan 一起工作了好几年，他很热情和并且主动性很高，主动编辑一本ceph的新用户综合指南。这个将会给部署开源社区版本 ceph 的用户带来有很大的帮助。

这本书包括了更多的技术文档以及社区的开发的一些新东西，给了新用户更有用的建议。

如果你下载了 ceph 的社区的版本，想体验它，想在家里部署它，或者在一个公司小环境上实现它，这个本书就是为你准备的。希望你能够根据建议，用例，来一步一步的部署集群，来验证集群的属性和功能。

现在，开始阅读这个本 Ceph Cookbook 并且开始部署你自己的软件定义存储。但更新的属性，例如生产版本的 cephfs 和容器的支持已经在筹备中，我们期待 karan 的下一本书。

Dr. Wolfgang Schulze

全球存储咨询公司总监，红帽

关于作者

Karan Singh 是一位IT专家和技术的传播者，与她的妻子生活在芬兰。他拥有学士学位，并且在计算机领域取得了 BITS 的硕士学位，除此之外，还取得了 OpenStack, NetApp, Oracle Solaris, Linux 等认证。

Karan 现在工作在云计算和存储领域，CSC-IT 数据中心，专注于开发基于 Openstack 和 Ceph 的云计算平台，并且使用 Ceph 构建了PB级别的存储系统。

Karan 拥有丰富的技能，并且在解决方案方面拥有很多的经验，云计算技术，自动化工具和 unix 系统。他也是ceph方面的第一本书 Learning Ceph 的作者。

Karan 用他的一部分时间用来研究和学习新的技术。当与 Ceph 和 Openstack 无关的工作的时候，他喜欢做一些新兴技术和自动化相关的工作。他喜欢写关于技术的博客。你可以在推特上 **@karansingh010** 找到他，或者使用 email 联系他 karan_singh1@live.com

> 我要感谢我的妻子，Monika 当我写书的时候为我准备了美味的食物， Kiitos MJ 你是一个伟大的厨师
>
> 我想借此机会感谢我的公司，CSC-IT 科学数据中心，以及所有与我共同工作过的同事，CSC-IT 是一个很好的工作的地方
>
> 我也想感谢 Ceph 社区充满活力的向前发展着，感谢 Ceph 的开发，改进和生态系统。
>
> 最后我要感谢整个出版团队，以及发行过程中的审阅者们

关于审稿

Christian Eichelmann 曾担任系统工程师，并且在德国担任了数年的IT架构师，经历过很多不同的公司，他一直在使用 CEPH，最早的版本还是早期的 alpha 版本，目前运行了多个PB级别的集群，他还开发了 ceph-dash：一个流行的监控 ceph 的图形界面。

Haruka Iwao 是一个 Google 的广告解决方案的工程师，她曾经作为一名存储解决方案架构师在红帽工作过，并且为ceph社区做出了自己的贡献，她还在日本的一些初创的公司里面担任过可靠性工程师，她对可靠性和可扩展性的计算比较感兴趣，他在 Tsukuba 大学里面的硕士课程就是分布式系统。

序论

我们是数字世界的一部分，每一秒都在产生着大量的数据。数据的增长速度是不可想象的，根据估算，人类在2020年将会产生40ZB的数据，也许它不是太多，但在2050年？我们估计有 Yottabyte(尧字节)数据？最明显的问题是：我们用什么方法来存储这么巨大的数据，我们准备好了么？对我来说，Ceph 就是这个希望，并且这个可能是未来十年都需要的存储技术。Ceph是存储的未来。

有人说：“软件走天下”，这是真的，从另外一个角度来看，软件是用可行的方法去计算各种需求，天气的计算，网络，存储，数据中心等，正如你所知道的，人们将自己的想法寄托在计算机上面，然后解决了很多难题，我认为，这些软件的方式将会解决未来计算的问题。

Ceph 是一个真正开源的，软件定义存储的解决方案，特意用性能线性增长的方式来解决数据增长的问题。它提供了文件，对象，块存储的统一存储方式，更多的属性包括，分布式，可扩展，可靠和自动化的架构，而且它是经济的存储方式，它能让你创造更多的价值。

Ceph 是存储行业的一个新亮点，它的企业级的属性包括，可扩展性，高可靠性，高速分层，纠删码和其它的，已经在过去几年取得了显著的改善，举几个例子，CERN，Yahoo,Dreamhost等，很多地方都部署了PB级别的集群，并且运行的很成功。

因为 Ceph 的块和对象存储已经开发和实施了一段时间了，直到去年，Cephfs 是生产环境唯一没有准备好的，这一年我将关注 Cephfs，在 Ceph Jewel版本将是生产版本，我迫不及待的想看到 Cephfs 在生产环境当中使用的例子，在几个领域 Ceph 得到了普及，例如AFA（All Flash Array），数据库，容器，虚拟化，好吧，Ceph 才刚刚开始，最好的即将到来。

在这本书中，我们将深入的了解 Ceph-包括它的组件和架构，还有怎么工作的。这本Ceph Cookbook 侧重于动手方面的知识，提供一步一步的指导。从第一章开始，按照章节一步步的操作，你将会获得实践经验，你将学习到一些有趣的概念。我希望通过这本书能够掌握 Ceph，在概念上和实践上，你将能够有信心并且能够成功的操作你的ceph存储集群。

快乐学习

Karan Singh

这本书将覆盖哪些内容

章节 1, Ceph – 介绍以及其它, 涵盖了介绍 Ceph，以及 RAID 和它的挑战。Ceph 的架构的概览。最后我们将安装和配置Ceph。

章节 2, Ceph Block Device 相关, 涵盖了 Ceph 块设备的介绍和配置.我们讲介绍 RBD 的快照，克隆，以及作为Openstack的cinder，glance和nova的配置。

章节 3,Ceph 对象存储相关，将深入了解ceph的对象存储，包括 RGW,联合网关设置，S3，以及 Openstack swift接口。最后我们将使用ownCloud配置同步和文件服务。

章节 4, Ceph 文件系统相关，包括 CephFS 的介绍，部署 MDS 和 Cephfs，通过内核客户端，Fuse 客户端，以及 NFS-Ganesha 访问。你还将学习到怎么通过 ceph-dokan 在 Windows 下面访问 CephFS。

章节 5, 使用 Calamari 监控 Ceph 集群, 包括通过CLI，介绍 Calamari，配置 Calamari 服务端和客户端。还将介绍通过 Calamari 的图形界面来监控集群，以及 Calamari 的故障排查。

章节 6, 操作和管理 Ceph 集群, 包括服务的管理，扩展和缩小集群。本章还包括故障硬盘的更换，以及升级Ceph 的基础设施。

章节 7, Ceph 的内部运作, 探索 Ceph 的 CRUSH Map,理解 CRUSH map的内部，随后将讲 Ceph 的认证和授权。本章还将讲述动态集群的管理，理解 Ceph 的PG.最后，我们讲述具体硬件需要了解的细节。

章节 8, 生产环境的规划和性能调优, 包括生产环境部署的规划。Ceph 的硬件和软件的规划。本章还包括Ceph 的建议和性能调优。最后，本章将涵盖纠删码和缓存分层。

章节 9, Ceph 的 virtual storage manager（intel的vsm）, 包括其介绍和体系架构，我们将部署VSM，然后使用VSM创建并管理Ceph。

章节 10, 更多关于 Ceph, 书的最后一章，包括 Ceph 的基准测试，使用 admin socket 和 API 进行故障调试. ceph-objectstore tool。本章还介绍使用 Ansible，Ceph 的内存分析。

你需要为这本书准备什么?

本书中所需要的软件如下：

- VirtualBox 4.0 or higher (https://www.virtualbox.org/wiki/Downloads)
- GIT (http://www.git-scm.com/downloads)
- Vagrant 1.5.0 or higher (https://www.vagrantup.com/downloads.html)
- CentOS operating system 7.0 or higher (http://wiki.centos.org/Download)
- Ceph software packages Version 0.87.0 or higher (http://ceph.com/resources/ downloads/) f
- S3 Client, typically S3cmd (http://s3tools.org/download)
- Python-swift client
- ownCloud 7.0.5 or higher (https://download.owncloud.org/download/ repositories/stable/owncloud/)
- NFS Ganesha
- Ceph Fuse

- Ceph-Dokan
- Ceph-Calamari (https:\/github.com\/ceph\/calamari.git)
- Diamond (https:\/github.com\/ceph\/Diamond.git)
- Ceph Calamari Client, romana (https:\/github.com\/ceph\/romana) Virtual Storage Manager 2.0 or higher (https:\/github.com\/01org\/virtualstoragemanager\/releases\/tag\/v2.1.0)
- Ansible 1.9 or higher (http:\/docs.ansible.com\/ansible\/intro_ installation.html)
- OpenStack RDO (http:\/rdo.fedorapeople.org\/rdo-release.rpm)

这本书给谁看的?

这本书的目标群体是存储和云计算工程师，系统管理员，IT 架构师，以及基于 CEPH 的软件定义存储的解决方案顾问，用来增强他们的云存储架构。如果你拥有 GNU\/Linux 和存储的基本知识，但没有软件定义存储和ceph相关的经验，但是渴望学习，这本书就是为你准备的。

段落结构

在这本书当中，你会发现会频繁的出现下面几个标题（准备工作，如果去做，它是如何工作的，更多的， 还可以参阅 ）

为了更清楚的完成这些步骤，我们使用下面的段落：

准备工作

本节将会告诉你这篇将会有什么期望，介绍如何配置一些软件或者一些基础的配置。

如何去做...

本节包括了详细的步骤

它是如何工作的...

本节将会详细解释上一节的工作原理

更多的...

这一部分将会讲述这一节内容的其它相关信息，以便读者掌握更多的相关知识

还可以参阅

本节提供了一些跟这一章节相关的有用的链接资源

约定

在这本书当中，你会发现用不同种类的字体来区分不同的信息。下面是这些字体风格的例子和它们含义的解释

文本中的变量名称，数据库表名，文件夹名，文件扩展名，路径名，虚拟的网址，用户输入，带@的名称，会是这样的例子：“为了做这个，我们需要编辑 OpenStack 节点的 `/etc/nova/nova.conf` 配置文件，根据下面的步骤添加内容”

一个代码块格式如下：

```
    inject_partition=-2
    images_type=rbd
    images_rbd_pool=vms
    images_rbd_ceph_conf=/etc/ceph/ceph.conf
```

命令行的输入或者输出是这样的：

```
# rados -p cache-pool ls
```

新的术语 和 重要的单词 会粗体显示。你现在看到的这行就是，例如，在菜单和对话框中，出现下面的提示：“定位到 **Optionsdefined** 的 **nova.virt.libvirt.volume** 的部分，增加下面的代码：”

> 警告或者重要的提示会这样的样式

读者反馈

我们欢迎读者进行反馈。让我们知道你是如何看待这本书的--你喜欢什么或者不喜欢什么。读者的反馈对我们很重要，可以帮助我们清楚怎样才能让书中知识更好的传递给读者。可以发送邮件到 feedback@packtpub.com 进行反馈，并注明书的标题。如果你在某些领域有专业的知识，想写一本书或者做一些书的发行相关工作，请访问我们的网站 www.packtpub.com/authors

客户支持

现在你拥有了 Packt book ，我们有很多事情帮助你从书中获得更多的价值

下载示例代码

你可以通过你的http://www.packtpub.com 的账户下载所有你购买过的书的示例代码。如果你在其它地方购买的这本书，你可以访问 http://www.packtpub.com/support 并且注册，相关文件可以通过邮件发送给你。

勘误表

尽管我们已经尽力确保我们的内容的准确性，但是错误难免会发生，如果你发现我们的一个错误--也许文字或者代码错误--我们很感激你能够发生给我们。这样可以减少后续读者的麻烦并且能够帮助我们提高这本书后续版本的质量。如果你发现了任何错误，请通过访问 http://www.packtpub.com/submit-errata ，选择你的书，点击提交勘误表的链接，一旦你的勘误表被核实，你的勘误表将被接受并上传到我们的网站或者加入到已经存在的勘误表当中。要查看之前提交的勘误表，访问 https://www.packtpub.com/books/ content/support ，在搜索框输入这本书名。所需要的信息将会出现在勘误部分

盗版

本书为内部使用的手册，禁止私下传播，以免不必要的版权争端

问题

如果你有这本书的任何方面的问题，您可以与我们联系 questions@packtpub.com，我们将尽我们所能解决问题

一、ceph - 介绍以及其它

在这个章节，我们将会覆盖下面的内容：

- Ceph - 一个新的时代开启
- RAID - 一个时代的结束
- Ceph - 架构的概述
- Ceph 部署规划
- 配置一个虚拟的环境
- 安装和配置 Ceph
- 扩展你的集群
- 集群操作实践

介绍

Ceph 是目前最热门的软件定义存储技术（**SDS**）,影响了整个存储行业。它是一个开源的项目，提供了块，文件和对象存储统一的存储解决方案。Ceph 的核心目标是提供一个分布式文件系统：可大规模扩展，高性能，没有单点故障。从底层来看，它已经被设计成高度可扩展，在通用商用硬件上部署，可以达到 EB 或者更高级别。

Ceph 之所以能获得存储行业这么大的关注度，归功于它的开放，可扩展和可靠的属性。现在是云计算和软件定义基础架构的时代，需要一个纯粹的软件定义存储的后端，更重要的是，云已经准备好了。无论您是运行的公有云，私有云或者混合云， Ceph 都非常的适合。

如今的软件都非常的智能，能够最大化的利用商用硬件来构建庞大的基础设施。Ceph 就是其中之一；它利用商用的硬件来提供企业级的自动化和高可靠存储系统。

Ceph 有着下面的架构理念使其不断的发展着：

- 每个组件都能够线性扩展
- 不能有任何的单点故障
- 解决方案需要是基于软件，开源，可适应的
- Ceph 软件能够运行在现有的通用的商业硬件上
- 每个组件都可以自我管理，自我修复的

Ceph 是基于对象存储之上的，这是它的基石，Ceph 这样的对象存储系统，满足了当前和未来非结构化数据对存储的需求。对象存储有其比传统存储解决方案更好的地方：可以实现存储平台和硬件的独立性。Ceph 控制对象分布到集群当中，并且复制它们，对象不依赖于物理路径，单独的标记对象的位置。这些灵活的属性，使得 Ceph 能够线性的从 PB 扩展到 EB 级别。

Ceph 提供强大的性能，高可扩展性和适应性。它可以帮助解决昂贵的存储孤岛问题。Ceph 确实是运行在通用商业软件上的企业级的存储解决方案；它是低成本，而功能丰富的存储系统。Ceph 在一套系统里面提供了块，文件和对象存储，让用户能够想用哪种就用哪种。

Ceph 版本

Ceph 正在以迅猛的速度开发着。2012年7月13日，Sage 发布了 Ceph 的第一个长期支持版本 Argonaut，从那时起，我们已经看到了七个新的版本发布。Ceph的版本分为长期支持版本（Long Term Support），开发者版本，每隔一个版本就是一个LTS版本。想要了解更多信息，访问 https://ceph.com/category/releases/

Ceph 版本名称	**Ceph**版本号	发布日期
Argonaut	V0.48 (LTS)	July 3, 2012
Bobtail	V0.56 (LTS)	January 1, 2013
Cuttlefish	V0.61	May 7, 2013
Dumpling	V0.67 (LTS)	August 14, 2013
Emperor	V0.72	November 9, 2013
Firefly	V0.80 (LTS	May 7, 2014
Giant	V0.87.1	Feb 26, 2015
Hammer	V0.94 (LTS)	April 7, 2015
Infernalis	V9.0.0	May 5, 2015
Jewel	V10.0.0	Nov, 2015

> 这里可以发现：ceph的版本名称是遵循字母表顺序的；下一个版本将会是 “K” 版本。
>
> "Ceph"的是给章鱼类的共同的昵称，这是多足类软体海洋动物的简称。Ceph 用章鱼作为其吉祥物，代表Ceph 具有高度并发的属性，类似于章鱼。

Ceph - 一个新的时代开启

数据对存储的需求在过去的几年里爆炸性的增长。研究表明，在大型的组织中，每年的数据量在以40%到60%的速度增长着，许多公司每年增长了一倍的数据量。IDC 的分析师估计，在2000年的时候，全球的总数据量为54.4 EB 。在2007年的时候，达到了 295 EB，到2020年的时候，有望达到 44 ZB 。这样的增长速度，传统存储是无法管理的；我们需要一个像 Ceph 这样的存储系统，它是分布式的，可扩展的，更重要的是经济上可行的系统。Ceph 已经特别设计，来满足当前以及未来对存储的需求。

软件定义存储（**SDS**）

SDS是用来降低存储基础设施的 TCO 的。除了降低存储的成本，一个SDS可以提供灵活性，可扩展性和可靠性。Ceph 是一个真正的 SDS 解决方案；它运行在通用的商用硬件之上，不担心于厂商硬件绑定，能够提供更低的 单GB价格。不同于传统的存储，软硬件绑为一体，在SDS中，你可以自由的从任何厂商购买硬件，可以随意的设计自己需要的异构的硬件解决方案。Ceph 的软件定义存储运行在硬件之上，提供很多自动化的处理，从软件层去提供很多企业存储的属性。

云存储

存储是云基础设施存在的弊端之一。每一个云基础设施都需要一套可靠的，低成本的，可扩展的存储，能够比其它的组件更紧密的结合。有很多传统的存储解决方案都自称已经准备好对接云存储了，但是我们今天不仅仅需要的是能够对接云存储，我们需要的是能够跟云系统完全的集成，可以提供更低的 TCO ,具备高可靠性和高可扩展性。云基础架构是建立在通用商业硬件之上的；同样的它需要存储系统也是建立在通用商业硬件之上的，Ceph 是云基础架构存储的最佳选择。

Ceph 迅速的发展着，已经离真正需要的云存储需求很近了。它已经取得多个主要开源云平台存储部分的中心地位，像 OpenStack，CloudStack 和 OpenNebula。此外Ceph 还成功的与云计算厂商 Red Hat, Canonical, Mirantis, SUSE,还有很多公司达成了合作伙伴关系。这些公司花了大量的时间来将 Ceph 作为其 OpenStack 发行版本的官方存储后端，这些都让 Ceph 在云存储行业变得非常热门。

OpenStack 是驱动公有云和私有云的最佳开源软件范例之一。它已经证明了自己是一个端到端的开源解决方案。 OpenStack 是一些软件的集合，例如 cinder,glance 和 swift，这些都是 OpenStack的存储组件。这些存储组件需要一个像Ceph这样的可靠的，可扩展的，全部在一起的存储后端。正因为如此，OpenStack和Ceph的社区在一起工作了很多年，来让 Ceph 成为一个完全兼容 OpenStack 的存储后端。

基于 Ceph 的云存储基础设施，提供了很大的灵活性去构建存储即服务（Storage-as-a-Service），基础架构即服务（ Infrastructure-as-a-Service）的解决方案，这是其它传统企业存储解决方案无法实现的，因为它们设计之初不是为了满足云计算的。通过使用 Ceph ，服务提供商能够为他们的客户提供低成本，高可靠的云存储服务。

下一代统一存储架构

统一存储的定义近些年发生了一些变化。几年前，所谓的 “统一存储”是指从单一的系统提供文件和块存储。现在，因为最近的科技的进步，例如云计算，大数据，物联网，一种新类型的存储一直在发展着，这就是对象存储。所有的不支持对象存储接口的系统不是真正的统一存储解决方案。一个真正的统一存储就应该像Ceph这样：能够从同一套存储系统中提供块，文件和对象存储。

在 Ceph 中，术语“统一存储”比现有的存储厂商所宣称的更有意义。Ceph 从一开始就是为未来的存储所设计，它的架构使得它能够处理大量的数据。当我们所说的“为未来做好准备”，是指它的对象存储功能，比文件和块存储更适合当今复杂混合的非结构化数据。不是直接管理块或者文件，Ceph 底层是存储的对象文件，然后基于对象之上提供块和文件存储。对象存储通过去除元数据的操作从而提供了更高的性能。Ceph 使用算法来计算和定位对象存储的位置。

传统的SAN和NAS系统的架构是非常受限的。基本的，他们遵循传统的高可用的处理方式，如果一个存储控制服务节点失效了，就从第二个节点提供服务。但是如果第二个节点也失效了，更糟糕的情况，所有的磁盘柜都发生了故障？大多数情况下，最终都是会丢失数据的。这种存储架构，不能承受多次的失效，这个不是我们现在需要的存储架构。传统存储架构的另外一个缺点就是它的数据存储和访问的机制。它拥有一个中心的元数据表来跟踪元数据，这意味着，每一次客户端发送的读写请求，存储系统需要先在巨大的元数据表中进行查找，在拿到了真正的数据位置以后，再接收客户端的请求。对于较小的存储系统来说，可能没有太多的影响性能，但是对于很大的集群来说-性能会受到很大的制约。同样会限制系统的可扩展性。

Ceph 没有遵循这样传统的存储架构；事实上，这个架构是全新设计的。没有去存储和操作元数据，Ceph引入了一个新的方式：CRUSH算法。CRUSH的意思是 Controlled Replication Under Scalable Hashing（可扩展散列下的可控制的复制）。不像以前的存储那样在客户端请求的时候需要去做元数据表查询，CRUSH算法计算出数据的写入和读取的路径。通过计算出的元数据，就不需要一个中心的元数据表了。当前的计算机的性能非常好，可以非常快的执行CRUSH查找，这个计算负载不会很高，并且可以利用分布式存储优点将计算分布到所有节点。除了这一点，CRUSH还有独特的属性，能够识别基础设施。能够识别存储单元之间的关系，将数据存储到正确的故障域，例如：磁盘，节点，机架，机架排，数据中心等等。通过 CRUSH 算法存储所有故障域当中，这样即使部分故障域失效了，还是可用的。正是因为有了 CRUSH，Ceph能够处理多重故障，提供高可靠和持久的存储。

CRUSH 算法使得 Ceph 能够自我管理和自我修复。在一个故障域当中某个部分失效的时候，CRUSH能够感知到。在没有任何人为的干预下，CRUSH 能够通过触发失效数据的恢复，自我管理和修复。CRUSH 会通过还有的数据重新生成丢失的数据。如果你正确的配置了 CRUSH map，它确保至少有一份数据是可以访问的。通过 CRUSH 我们能设计出一个无单点故障的高可用的存储基础设施。这让 Ceph 成为一个高度可扩展和高可靠的能适应未来的存储。

RAID - 一个时代的结束

RAID技术已经成为块存储系统的基本组成很多年了。在过去的三十年中，RAID 几乎适应所有类型数据，但是，所有的时代都会走到了尽头，而这一次，轮到RAID了。这些系统已经开始显现出其局限性，没有能力满足未来的存储需求。在过去的几年过程中，云基础架构已经取得了飞速的发展，产生了很多新的存储需求，这个对于传统的RAID系统来说是个新的挑战。在本节中，我们会介绍RAID系统的局限性。

RAID 重建是痛苦的

在 RAID 技术中最痛苦的是它超级漫长的重建过程。磁盘生产厂商正在增加单个磁盘的容量，每 GB 磁盘的价格越来越低。我们不再谈论450 GB，600 GB，甚至1 TB磁盘，今天有更大容量的磁盘。较新的企业级磁盘硬盘提供的容量高达 4TB，6TB，甚至10TB。并且磁盘容量保持逐年增加。想一想，如果一个由无数的 4TB 或者 6TB 磁盘组成的企业级RAID存储系统。如果不走运的话，当其中某个磁盘驱动器发生故障，RAID将需要几个小时，甚至多达几天修复单个磁盘的故障。如果在恢复的同时， 同一个RAID组的另一个磁盘出现故障，那么将会是一个混乱的局面。RAID修复多个大容量磁盘将会是一个繁琐的过程。

RAID备用磁盘增加TCO

RAID系统需要几个磁盘作为热备盘。这些都是空闲磁盘，只有在一个磁盘发生故障时才会使用，否则它们不会被用于存储数据。这增加了额外的系统成本，增加了TCO（总体拥有成本）。此外，如果你不使用备用磁盘，一旦RAID组中磁盘失效，那么你将面临比较严重的问题。

RAID 比较贵，并且依赖于硬件

RAID组需要一组相同的磁盘；如果更改磁盘大小，转速，或磁盘类型，将会对存储系统的容量和性能产生不利影响。这使得RAID在硬件选择上非常挑剔。

此外，企业级RAID存储系统通常需要昂贵的硬件组件，例如RAID控制器，这显然增加了系统的成本。 如果你没有很多这样的系统，RAID控制器可能发生单点故障。

RAID组的扩容是一个挑战

当RAID组不能够扩容的时候将是一个比较窘迫的场景，因为RAID不支持横向扩展，超过一个节点的时候，将不能跨节点做RAID，即使你有再多的钱也不行。一些系统支持增加磁盘架，但是这个能够增加的容量也是非常有限的；并且这些新的磁盘的负载将会增加到原来的存储控制器上。所以你需要在容量和性能之间做一个权衡。

RAID可靠性模型已经不被看好

RAID可以配置成多种不同的类型；最常见的类型是RAID5和RAID6，其可以分别允许一个和两个磁盘的失效。RAID在两个磁盘故障后就不能确保数据的可靠性。这是RAID系统的最大缺点之一。

此外，在一个RAID重建过程中，客户端的IO请求将会受到非常大的影响，直到重建完成才能恢复。RAID的另一个限制因素是，它只能防止磁盘故障；它不能防止网络，硬件，操作系统，电源等的故障以及其它的数据中心内的灾难。

讨论完RAID的弊端之后，我们可以得出这样的结论：我们现在需要一个能够克服性能缺点，能够控制好成本的系统。Ceph存储系统是当今解决这些问题的最佳解决方案之一。让我们看看Ceph如何解决的。

为保证可靠性，Ceph利用数据复制的方法，这意味着它不使用RAID，从而克服一切基于RAID的企业级系统中发现的问题。Ceph的是一个软件定义存储，所以我们并不需要专用的硬件；此外，副本的级别是高度可自定义的，这意味着Ceph的存储管理员可以配置一个最小副本为1，或者很高的副本数量的系统，这完全取决于底层基础架构。

在一个或多个磁盘故障的情况下，Ceph的复制是一个比RAID更好的过程。当一个磁盘出现故障，在那个时间点磁盘上所拥有的数据，开始从它的对等磁盘上进行恢复。因为 Ceph 是一个分布式系统，所有的数据副本都是对象的形式，分散在整个群集上的磁盘上，定义一个不同的故障域，使得不会出现两个对象的副本在同一磁盘上。这个比较好的情况是所有集群磁盘都参加数据恢复。这使得恢复操作非常的快，从而性能问题很少。更好的情况是，恢复操作不需要任何备用磁盘；数据简单地复制到集群中的其它Ceph磁盘。Ceph的磁盘有一个权重的机制，从而不同的磁盘大小也是没有问题的。

除了副本方法，Ceph 还支持另一种先进的数据可靠性的方式：使用纠删编码技术。相比于副本池，纠删池需要更少的存储空间。在纠删码中，数据通过纠删编码计算出恢复的数据。在同一个集群中，副本和纠删都可以使用，只是对应到不同的存储池。在接下来的章节中我们将详细了解纠删编码技术。

二、**Ceph Block Device** 相关