Nick Fisk

Mastering Ceph

Redefine your storage system

Packt>

Mastering Ceph

Mastering Ceph covers all that you need to know to use Ceph effectively. Starting with design goals and planning steps that should be undertaken to ensure successful deployment, you will be guided through setting up and deploying the Ceph cluster with the help of orchestration tools. Key areas of Ceph including Bluestore, erasure coding, and cache tiering will be covered with the help of examples. The development of applications that use Librados and distributed computations with shared object classes are also covered. A section on tuning will take you through the process of optimisizing both Ceph and its supporting infrastructure. Finally, you will learn how to troubleshoot issues and handle various scenarios where Ceph is likely not to recover on its own.

By the end of the book, you will be able to successfully deploy and operate a resilient, high performance Ceph cluster.

Things you will learn:

- Know when and how to use some of Ceph's advanced new features
- Set up a test cluster with Ansible and some virtual machines using VirtualBox and Vagrant
- Develop novel solutions to massive problems with librados and shared object classes
- Choose intelligent parameters for an erasure coded pool and set it up
- Configure the Bluestore settings and see how they interact with different hardware configurations
- Keep Ceph running through thick and thin with tuning, monitoring, and disaster recovery advice

Packt
www.packtpub.com

$ **44.99** US
£ **37.99** UK

Prices do not include local sales
Tax or VAT where applicable

ISBN 978-1-78588-878-6
54499
9 781785 888786

Mastering Ceph

Nick Fisk

Mastering Ceph

Redefine your storage system

Nick Fisk

Packt

BIRMINGHAM - MUMBAI

Mastering Ceph

Copyright © 2017 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor Packt Publishing, and its dealers and distributors will be held liable for any damages caused or alleged to be caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

First published: May 2017

Production reference: 1260517

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham
B3 2PB, UK.

ISBN 978-1-78588-878-6

www.packtpub.com

此为试读，需要完整PDF请访问：www.ertongbook.com

Credits

Author
Nick Fisk

Reviewer
Vladimir Franciz S. Blando

Commissioning Editor
Pratik Shah

Acquisition Editor
Meeta Rajani

Content Development Editor
Juliana Nair

Technical Editor
Gaurav Suri

Copy Editor
Dipti Mankame

Project Coordinator
Judie Jose

Proofreader
Safis Editing

Indexer
Rekha Nair

Graphics
Kirk D'Penha

Production Coordinator
Aparna Bhagat

About the Author

Nick Fisk is an IT specialist with a strong history in enterprise storage. Having worked in a variety of roles throughout his career, he has encountered a wide variety of technologies. In 2012, Nick was given the opportunity to focus more toward open source technologies, and this is when his first exposure to Ceph happened. Having seen the potential of Ceph as a storage platform and the benefits of moving away from the traditional closed-stack storage platforms, Nick pursued Ceph with a keen interest.

Throughout the following years, his experience with Ceph increased with the deployment of several clusters and enabled him to spend time in the Ceph community, helping others and improving certain areas of Ceph.

I would firstly like to thank my wife for allowing me to dedicate time to writing this book and to Juliana Nair for supporting me through the writing process.
I would also like to especially thank Tarquin Dunn for the continual support and encouragement he has given me from the initial adoption of Ceph as a storage platform. His support has enabled me to both learn and contribute to this amazing open source project.

About the Reviewer

Vladimir Franciz S. Blando is a seasoned IT professional with 18 years' experience of Linux systems administration, including architecture, design, installation, configuration, and maintenance, working on both bare-metal and virtualized environments. He is well versed in Amazon Web Services, OpenStack cloud, Ceph storage, and other cloud technologies.

I would like to thank Morphlabs for giving me the opportunity to work on various cloud technologies, especially OpenStack and Ceph, and for supporting me all throughout.

www.PacktPub.com

For support files and downloads related to your book, please visit www.PacktPub.com.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.PacktPub.comand as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at service@packtpub.com for more details.

At www.PacktPub.com, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.

https://www.packtpub.com/mapt

Get the most in-demand software skills with Mapt. Mapt gives you full access to all Packt books and video courses, as well as industry-leading tools to help you plan your personal development and advance your career.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Customer Feedback

Thanks for purchasing this Packt book. At Packt, quality is at the heart of our editorial process. To help us improve, please leave us an honest review on this book's Amazon page at `https://www.amazon.com/dp/1785888781`.

If you'd like to join our team of regular reviewers, you can e-mail us at `customerreviews@packtpub.com`. We award our regular reviewers with free eBooks and videos in exchange for their valuable feedback. Help us be relentless in improving our products!

此为试读,需要完整PDF请访问: www.ertongbook.com

Table of Contents

Preface 1

Chapter 1: Planning for Ceph 9

- **What is Ceph?** 9
- **How Ceph works?** 10
- **Ceph use cases** 11
  - Replacing your storage array with Ceph 11
  - Performance 11
  - Reliability 12
  - The use of commodity hardware 12
- **Specific use cases** 12
  - OpenStack- or KVM-based virtualization 13
  - Large bulk block storage 13
    - Object storage 13
    - Object storage with custom application 14
    - Distributed filesystem – web farm 14
    - Distributed filesystem – SMB file server replacement 14
- **Infrastructure design** 14
  - SSDs 14
    - Consumer 15
    - Prosumer 15
    - Enterprise SSDs 16
      - Enterprise – read intensive 16
      - Enterprise – general usage 16
      - Enterprise – write intensive 17
  - Memory 17
  - CPU 18
  - Disks 20
  - Networking 21
  - 10G networking requirement 21
- **Network design** 22
  - OSD node sizes 24
    - Failure domains 24
  - Price 25
  - Power supplies 26
- **How to plan a successful Ceph implementation** 26
  - Understanding your requirements and how it relates to Ceph 26

Defining goals so that you can gauge if the project is a success 27
Choosing your hardware 27
Training yourself and your team to use Ceph 28
Running PoC to determine if Ceph has met the requirements 28
Following best practices to deploy your cluster 29
Defining a change management process 29
Creating a backup and recovery plan 30
Summary 30
Chapter 2: Deploying Ceph 31
Preparing your environment with Vagrant and VirtualBox 32
System requirements 32
Obtaining and installing VirtualBox 33
Setting up Vagrant 34
The ceph-deploy tool 38
Orchestration 38
Ansible 39
Installing Ansible 40
Creating your inventory file 41
Variables 43
Testing 44
A very simple playbook 44
Adding the Ceph Ansible modules 46
Deploying a test cluster with Ansible 48
Change and configuration management 51
Summary 51
Chapter 3: BlueStore 53
What is BlueStore? 53
Why was it needed? 54
Ceph's requirements 54
Filestore limitations 54
Why is BlueStore the solution? 56
How BlueStore works 57
RocksDB 57
Deferred writes 58
BlueFS 58
How to use BlueStore 58
Upgrading an OSD in your test cluster 59
Summary 61

Chapter 4: Erasure Coding for Better Storage Efficiency 63
What is erasure coding? 64
K+M 64
How does erasure coding work in Ceph? 65
Algorithms and profiles 67
Jerasure 67
ISA 67
LRC 67
SHEC 68
Where can I use erasure coding? 68
Creating an erasure-coded pool 69
Overwrites on erasure code pools with Kraken 72
Demonstration 73
Troubleshooting the 2147483647 error 75
Reproducing the problem 76
Summary 78
Chapter 5: Developing with Librados 79
What is librados? 79
How to use librados? 80
Example librados application 83
Example of the librados application with atomic operations 88
Example of the librados application that uses watchers and notifiers 91
Summary 95
Chapter 6: Distributed Computation with Ceph RADOS Classes 97
Example applications and the benefits of using RADOS classes 97
Writing a simple RADOS class in Lua 98
Writing a RADOS class that simulates distributed computing 100
Preparing the build environment 100
RADOS class 102
Client librados applications 104
Calculating MD5 on the client 104
Calculating MD5 on the OSD via RADOS class 106
Testing 108
RADOS class caveats 110
Summary 111
Chapter 7: Monitoring Ceph 113
Why it is important to monitor Ceph 113
What should be monitored 114

Ceph health 114
Operating system and hardware 114
Smart stats 115
Network 115
Performance counters 115
PG states – the good, the bad, and the ugly 116
The good 116
The active state 116
The clean state 116
Scrubbing and deep scrubbing 116
The bad 116
The inconsistent state 117
The backfilling, backfill_wait, recovering, recovery_wait states 117
The degraded state 117
Remapped 117
The ugly 117
The incomplete state 117
The down state 118
The backfill_toofull state 118
Monitoring Ceph with collectd 118
Graphite 118
Grafana 121
collectd 123
Deploying collectd with Ansible 127
Sample Graphite queries for Ceph 129
Number of Up and In OSDs 130
Showing most deviant OSD usage 130
Total number of IOPs across all OSDs 130
Total MBps across all OSDs 130
Cluster capacity and usage 131
Average latency 131
Custom Ceph collectd plugins 131
Summary 133
Chapter 8: Tiering with Ceph 135
Tiering versus caching 135
How Cephs tiering functionality works 136
What is a bloom filter 137
Tiering modes 137
Writeback 137
Forward 138
Read-forward 138
Proxy 138

Read-proxy 138
Uses cases 138
Creating tiers in Ceph 139
Tuning tiering 142
Flushing and eviction 142
Promotions 145
Promotion throttling 146
Monitoring parameters 147
Tiering with erasure-coded pools 148
Alternative caching mechanisms 149
Summary 150
Chapter 9: Tuning Ceph 151
Latency 152
Benchmarking 154
Benchmarking tools 155
Fio 155
Sysbench 155
Ping 155
iPerf 155
Network benchmarking 155
Disk benchmarking 157
RADOS benchmarking 159
RBD benchmarking 160
Recommended tunings 161
CPU 161
Filestore 164
VFS cache pressure 164
WBThrottle and/or nr_requests 165
Filestore queue throttling 166
filestore_queue_low_threshhold 166
filestore_queue_high_threshhold 166
filestore_expected_throughput_ops 166
filestore_queue_high_delay_multiple 167
filestore_queue_max_delay_multiple 167
PG Splitting 167
Scrubbing 168
OP priorities 169
The Network 169
General system tuning 170
Kernel RBD 170
Queue Depth 171
ReadAhead 171

PG distributions 172
Summary 173
Chapter 10: Troubleshooting 175
Repairing inconsistent objects 176
Full OSDs 179
Ceph logging 180
Slow performance 181
Causes 181
Increased client workload 181
Down OSDs 181
Recovery and backfilling 182
Scrubbing 182
Snaptrimming 182
Hardware or driver issues 182
Monitoring 183
iostat 183
htop 183
atop 184
Diagnostics 184
Extremely slow performance or no IO 186
Flapping OSDs 186
Jumbo frames 186
Failing disks 186
Slow OSDs 187
Investigating PGs in a down state 187
Large monitor databases 188
Summary 189
Chapter 11: Disaster Recovery 191
What is a disaster? 192
Avoiding data loss 192
What can cause an outage or data loss? 192
RBD mirroring 193
The journal 194
The rbd-mirror daemon 194
Configuring RBD mirroring 194
Performing RBD failover 199
RBD recovery 201
Lost objects and inactive PGs 204
Recovering from a complete monitor failure 208
Using the Cephs object store tool 211

此为试读，需要完整PDF请访问：www.ertongbook.com