In this package, you will find:

- The author biography
- A preview chapter from the book, Chapter 1 “Introducing Ceph Storage”
- A synopsis of the book’s content
- More information on Learning Ceph

About the Author

Karan Singh is a curious IT expert and an overall tech enthusiast living with his beautiful wife, Monika, in Espoo, Finland. He holds a bachelor's (honors) degree in computer science and a master's degree in system engineering from BITS Pilani, India. In addition to this, he is a certified professional for technologies such as OpenStack, NetApp, and Oracle Solaris.

Karan is currently working as a system specialist of storage and platform for CSC – IT Center for Science Ltd. in Finland. He is actively involved in providing IaaS cloud solutions based on OpenStack and Ceph Storage at his workplace and has been building economic multipetabyte storage solutions using Ceph. Karan possesses extensive system administration skills and has excellent working experience on a variety of Unix environments, backup, enterprise storage systems, and cloud platforms.

When not working on Ceph and OpenStack, Karan can be found working with technologies such as Ansible, Docker, Hadoop, IoT, and other cloud-related areas. He aims to get a PhD in cloud computing and big data and wants to learn more about these technologies. He is an avid blogger at http://karan-mj.blogspot.fi/. You can reach him on Twitter as @karansingh010 and Ceph and OpenStack IRC channels as ksingh. You can also e-mail him at karan_singh1@live.com.
I’d like to thank my wife, Monika, for providing encouragement and patience throughout the writing of this book.

In addition, I would like to thank my company, CSC-IT Center for Science Ltd., and my colleagues for giving me an opportunity to work on Ceph and other cloud-related areas. Without CSC and Ceph, the opportunity to write this book would never have been possible. A big thanks goes out to the Ceph community for developing, improving, and supporting Ceph, which is an amazing piece of software.
Learning Ceph

Data—it's a simple word that stores the past, present, and future of the entire universe, and it's the most critical element of any system that exists today. We live in an era of technology that is generating enormous amount of data each second, and with time, this data growth will be unimaginable. However, how do we store this ever-growing data such that it remains secure, reliable, and future ready? This book is about one of the storage technology game-changers that will redefine the future of storage.

Ceph is an open source technology that is leading the way in providing software-defined storage. Ceph is an excellent package of reliability, unified nature, and robustness. It is more complete and economic than any other storage solution present today. Ceph has developed its own entire new way of storing data; it's distributed, massively scalable, with no single point of failure, and the best part is that it runs on commodity hardware, which makes it amazingly economic. It will give you power to break the knots of expensive vendor lock-in solutions and switch to enterprise-grade open source technology for all your storage needs.

Ceph has been enriched with enterprise-class features such as high degree of reliability, robustness, scalability, erasure coding, cache tiering, and many more. The maturity that Ceph has gained over the period of a decade makes it stand out in a crowd and lead the way to storage. It is the technology of today and the future; unified Ceph storage system is the solution for whatever requirements you have for data storage. Ceph is truly unified, that is, it has files, blocks, and objects in a single storage system; this makes Ceph extremely flexible and adaptable to all your data needs. It is the answer to all the data storage problems you have.

Cloud computing is the next paradigm shift, and a storage system such as Ceph is the most essential component of the cloud infrastructure. Ceph has its big footprint in the cloud storage area. It has been a leading open source software-defined storage choice for cloud platforms such as OpenStack and CloudStack. These cloud platforms leverage the features of Ceph and deliver robust, scalable, and exabyte-level public or private cloud infrastructures. In addition to this, virtualization platforms such as KVM and libvirt support Ceph big time, and support from proprietary virtualization solutions such as VMWare and HyperV is on the way.

Ceph is surely the next big thing in the storage industry, which is backed by Inktank, now a part of Red Hat. Ceph has an amazing community presence and quick development cycles, making it more reliable a couple of times a year. Though Ceph is purely open source, one can enjoy enterprise-support subscriptions from Red Hat and their partner ecosystem, which is an advantage.
What This Book Covers

*Chapter 1, Introducing Ceph Storage*, covers the evolution, history, as well as the future of Ceph. This chapter explains common storage challenges and how Ceph deals with them and becomes the game-changer. It also covers a comparison of Ceph with other storage systems.

*Chapter 2, Ceph Instant Deployment*, covers instant, step-by-step practical approaches to deploy your first Ceph cluster. It includes a guided tour of creating the Ceph sandbox environment on VirtualBox as well as scaling it up.

*Chapter 3, Ceph Architecture and Components*, dives deep into the Ceph internal architecture, explaining each and every component in detail. Components are explained sequentially and practically for greater learning and correlation.

*Chapter 4, Ceph Internals*, covers how Ceph manages data; the practical content will help you understand every piece of it. It also covers CRUSH, placement groups, and pools in detail.

*Chapter 5, Deploying Ceph – the Way You Should Know*, covers hardware planning required for a production-grade Ceph cluster. It also includes practical approaches of building a Ceph cluster in both the manual and automated ways using ceph-deploy.

*Chapter 6, Storage Provisioning with Ceph*, includes practical hands-on approaches to explain files, blocks, and object type storage in Ceph and how to configure and provision each type. The chapter also covers snapshots, cloning, S3- and swift-compatible object storage, and much more.

*Chapter 7, Ceph Operations and Maintenance*, covers everything to manage and operate Ceph from a system admin point of view. It includes daily operations, scaling up and down, hardware replacement, and a detailed coverage on CRUSH management and its advanced concepts.

*Chapter 8, Monitoring Your Ceph Cluster*, makes you competent in monitoring your Ceph cluster and all of its components. It also covers open source Ceph monitoring dashboard projects such as Kraken and ceph-dash and their installation and configuration.

*Chapter 9, Integrating Ceph with OpenStack*, covers step-by-step practical approaches to set up your own test OpenStack environment and integrating it with Ceph. It explains how Ceph benefits OpenStack and how OpenStack components make use of Ceph.

*Chapter 10, Ceph Performance Tuning and Benchmarking*, covers advanced concepts of Ceph, such as performance tuning from both hardware and software points of view. It also includes hands-on approaches to erasure coding and cache tiering and discusses Ceph benchmarking tools.
Introducing Ceph Storage

In this chapter, we will cover the following topics:

- An overview of Ceph
- The history and evolution of Ceph
- Ceph and the future of storage
- The compatibility portfolio
- Ceph versus other storage solutions

An overview of Ceph

Ceph is an open source project, which provides software-defined, unified storage solutions. Ceph is a distributed storage system which is massively scalable and high-performing without any single point of failure. From the roots, it has been designed to be highly scalable, up to exabyte level and beyond while running on general-purpose commodity hardware.

Ceph is getting most of the buzz in the storage industry due to its open, scalable, and distributed nature. Today, public, private, and hybrid cloud models are the dominant strategies for the purpose of providing massive infrastructure, and Ceph is getting popular in becoming a cloud storage solution. Commodity hardware is what the cloud is dependent on, and Ceph makes the best use of this commodity hardware to provide you with an enterprise-grade, robust, and highly reliable storage system.
Introducing Ceph Storage

Ceph has been raised and nourished with an architectural philosophy which includes the following features:

- Every component must be scalable
- There can be no single point of failure
- The solution must be software-based, open source, and adaptable
- Ceph software should run on readily available commodity hardware
- Everything must self-manageable wherever possible

Ceph provides great performance, limitless scalability, power, and flexibility to enterprises, thereby helping them get rid of expensive proprietary storage silos. Ceph is an enterprise-class, software-defined, unified storage solution that runs on commodity hardware, which makes it the most cost-effective and feature-rich storage system. The Ceph universal storage system provides block, file, and object storage under one hood, enabling customers to use storage as they want.

The foundation of Ceph lies on objects, which are its building blocks. Any format of data, whether it's a block, object, or file, gets stored in the form of objects inside the placement group of a Ceph cluster. Object storage such as Ceph is the answer for today's as well as the future's unstructured data storage needs. An object-based storage system has its advantages over traditional file-based storage solutions; we can achieve platform and hardware independence using object storage. Ceph plays intelligently with objects, and replicates each object across clusters to improve reliability. In Ceph, objects are not tied to a physical path, making objects flexible and location-independent. This enables Ceph to scale linearly from the petabyte level to an exabyte level.

The history and evolution of Ceph

Ceph was developed at University of California, Santa Cruz, by Sage Weil in 2003 as a part of his PhD project. The initial project prototype was the Ceph filesystem, written in approximately 40,000 lines of C++ code, which was made open source in 2006 under a Lesser GNU Public License (LGPL) to serve as a reference implementation and research platform. Lawrence Livermore National Laboratory supported Sage's initial research work. The period from 2003 to 2007 was the research period of Ceph. By this time, its core components were emerging, and the community contribution to the project had begun at pace. Ceph does not follow a dual licensing model, and has no enterprise-only feature set.
In late 2007, Ceph was getting mature and was waiting to get incubated. At this point, DreamHost, a Los-Angeles-based web hosting and domain registrar company entered the picture. DreamHost incubated Ceph from 2007 to 2011. During this period, Ceph was gaining its shape; the existing components were made more stable and reliable, various new features were implemented, and future roadmaps were designed. Here, the Ceph project became bona fide with enterprise options and roadmaps. During this time, several developers started contributing to the Ceph project; some of them were Yehuda Sadeh, Weinraub, Gregory Farnum, Josh Durgin, Samuel Just, Wido den Hollander, and Loïc Dachary, who joined the Ceph bandwagon.

In April 2012, Sage Weil founded a new company, Inktank, which was funded by DreamHost. Inktank was formed to enable the widespread adoption of Ceph's professional services and support. Inktank is the company behind Ceph whose main objective is to provide expertise, processes, tools, and support to their enterprise-subscription customers, enabling them to effectively adopt and manage Ceph storage systems. Sage was the CTO and Founder of Inktank. In 2013, Inktank raised $13.5 million in funding. On April 30, 2014, Red Hat, Inc.—the world’s leading provider of open source solutions—agreed to acquire Inktank for approximately $175 million in cash. Some of the customers of Inktank include Cisco, CERN, and Deutsche Telekom, and its partners include Dell and Alcatel-Lucent, all of which will now become the customers and partners of Red Hat for Ceph's software-defined storage solution. For more information, please visit www.inktank.com.

The term Ceph is a common nickname given to pet octopuses; Ceph can be considered as a short form for Cephalopod, which belongs to the mollusk family of marine animals. Ceph has octopuses as its mascot, which represents Ceph's parallel behavior to octopuses.

The word Inktank is somewhat related to cephalopods. Fishermen sometimes refer to cephalopods as inkfish due to their ability to squirt ink. This explains how cephalopods (Ceph) have some relation with inkfish (Inktank). Likewise, Ceph and Inktank have a lot of things in common. You can consider Inktank to be a thinktank for Ceph.

Sage Weil is one of the cofounders of DreamHost.
Introducing Ceph Storage

Ceph releases
During late 2007, when the Ceph project started, it was first incubated at DreamHost. On May 7, 2008, Sage released Ceph v0.2, and after this, its development stages evolved quickly. The time between new releases became short and Ceph now has new version updates every next month. On July 3, 2012, Sage announced a major release with the code name Argonaut (v0.48). The following are the major releases of Ceph, including Long Term Support (LTS) releases. For more information, please visit https://ceph.com/category/releases/.

<table>
<thead>
<tr>
<th>Ceph release name</th>
<th>Ceph release version</th>
<th>Released in</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argonaut</td>
<td>v0.48 (LTS)</td>
<td>July 3, 2012</td>
</tr>
<tr>
<td>Bobtail</td>
<td>v0.56 (LTS)</td>
<td>January 1, 2013</td>
</tr>
<tr>
<td>Cuttlefish</td>
<td>v0.61</td>
<td>May 7, 2013</td>
</tr>
<tr>
<td>Dumpling</td>
<td>v0.67 (LTS)</td>
<td>August 14, 2013</td>
</tr>
<tr>
<td>Emperor</td>
<td>v0.72</td>
<td>November 9, 2013</td>
</tr>
<tr>
<td>Firefly</td>
<td>v0.80 (LTS)</td>
<td>May 2014</td>
</tr>
<tr>
<td>Giant</td>
<td>v0.87</td>
<td>(Future release)</td>
</tr>
</tbody>
</table>

Ceph release names follow alphabetical order; the next release will be named with the initial I.

Ceph and the future of storage
Enterprise storage requirements have grown explosively over the last few years. Research has shown that data in large enterprises is growing at a rate of 40 to 60 percent annually, and many companies are doubling their data footprint each year. IDC analysts estimated that there were 54.4 exabytes of total digital data worldwide in the year 2000. By 2007, this reached 295 exabytes, and by the end of 2014, it’s expected to reach 8,591 exabytes worldwide.

Worldwide storage demands a system that is unified, distributed, reliable, high performance, and most importantly, massively scalable up to the exabyte level and beyond. The Ceph storage system is a true solution for the growing data explosion of this planet. The reason why Ceph is emerging at lightning pace is its lively community and users who truly believe in the power of Ceph. Data generation is a never-ending process. We cannot stop data generation, but we need to bridge the gap between data generation and data storage.
Ceph fits exactly in this gap; its unified, distributed, cost-effective, and scalable nature is the potential solution to today’s and the future’s data storage needs. The open source Linux community had foreseen Ceph’s potential long back in 2008, and they had added support for Ceph in the mainline Linux kernel. This has been a milestone for Ceph as there is no other competitor to join it there.

Ceph as a cloud storage solution
One of the most problematic areas in cloud infrastructure development is storage. A cloud environment needs storage that can scale up and out at low cost and which can be easily integrated with other components of that cloud framework. The need of such a storage system is a vital aspect to decide the total cost of ownership (TCO) of the entire cloud project. There are several traditional storage vendors who claim to provide integration to the cloud framework, but today, we need additional features beyond just integration support. These traditional storage solutions might have proven successful a few years back, but at present, they are not a good candidate for being a unified cloud storage solution. Also, traditional storage systems are too expensive to deploy and support in the long run, and scaling up and out is a gray area for them. Today, we need a storage solution that has been totally redefined to fulfill the current and future needs, a system that has been built upon open source software, and commodity hardware that can provide the required scalability in a cost-effective way.
Introducing Ceph Storage

Ceph has been rapidly evolving in this space to bridge this gap of a true cloud storage backend. It is grabbing center stage with every major open source cloud platform such as OpenStack, CloudStack, and OpenNebula. In addition to this, Ceph has built partnerships with Canonical, Red Hat, and SUSE, the giants in Linux space. These companies are favoring big time to Ceph—the distributed, reliable, and scalable storage clusters for their Linux and cloud software distributions. Ceph is working closely with these Linux giants to provide a reliable multifeatured storage backend for their cloud platforms.

Public and private clouds are gaining a lot of momentum due to the OpenStack project. OpenStack has proven itself as an end-to-end cloud solution. It has its internal core storage components named Swift, which provides object-based storage, and Nova-Volume, also known as Cinder, which provides block storage volumes to VMs.

Unlike Swift, which is limited only to object storage, Ceph is a unified storage solution of block, file, and object storage, and thus benefits OpenStack by providing multiple storage types from a single storage cluster. So, you can easily and efficiently manage storage for your OpenStack cloud. The OpenStack and Ceph communities have been working together for many years to develop a fully supported Ceph storage backend for the OpenStack cloud. Starting with Folsom, which is the sixth major release of OpenStack, Ceph has been fully integrated with it. The Ceph developers ensured that Ceph works well with the latest version of OpenStack, and at the same time, contribute to new features as well as bug fixes. OpenStack utilizes one of the most demanding feature of Ceph, the **RADOS block device (RBD)**, through its cinder and glance components. Ceph RBD helps OpenStack in rapid provisioning of hundreds of virtual machine instances by providing snapshotted-cloned volume, which are thin-provisioned, and hence less space hungry and ultra quick.

Cloud platforms with Ceph as a storage backend provide the much needed flexibility to service providers to build Storage-as-a-Service and Infrastructure-as-a-Service solutions, which they cannot achieve from other traditional enterprise storage solutions, as they are not designed to fulfill cloud needs. Using Ceph as a backend for cloud platforms, service providers can offer low-cost cloud services to their customers. Ceph enables them to offer relatively low storage prices with enterprise features compared to other storage providers such as Amazon.

Dell, SUSE, and Canonical offer and support deployment and configuration management tools such as Dell Crowbar and Juju for automated and easy deployment of Ceph storage for their OpenStack cloud solutions. Other configuration management tools such as Puppet, Chef, SaltStack, and Ansible are quite popular for automated Ceph deployment. Each of these tools has its open source, readymade Ceph modules that can be easily used for Ceph deployment. In a distributed environment such as Cloud, every component must scale. These configuration management tools are essential to quickly scale up your infrastructure.
Ceph is now fully compatible with these tools, allowing customers to deploy and extend a Ceph cluster instantly.

Starting with the OpenStack Folsom release, the nova-volume component has become cinder; however, nova-volume commands still work with OpenStack.

Ceph as a software-defined solution
All the customers who want to save money on storage infrastructure are most likely to consider **Software-defined Storage (SDS)** very soon. An SDS can offer a good solution to customers with a large investment in legacy storage who are still not getting required flexibility and scalability. Ceph is a true SDS solution, which is an open source software, runs on any commodity hardware, hence no vendor lock in, and provides low cost per GB. An SDS solution provides the much needed flexibility with respect to hardware selection. Customers can choose any commodity hardware from any manufacturer and are free to design a heterogeneous hardware solution for their own needs. Ceph's software-defined storage on top of this hardware will take care of everything. It also provides all the enterprise storage features right from the software layer. Low cost, reliability, and scalability are its main traits.

Ceph as a unified storage solution
The definition of a unified storage solution from a storage vendor's perspective is comprised of file-based and block-based access from a single platform. The enterprise storage environment provides NAS plus SAN from a single platform, which is treated as a unified storage solution. NAS and SAN technologies were proven to be successful in the late 90’s and early 20’s, but if we think about the future, are we sure that NAS and SAN can manage storage needs 50 years down the line? Do they have enough potential to handle multiexabytes of data? Probably not.

In Ceph, the term **unified storage** is more meaningful than what existing storage vendors claim to provide. Ceph has been designed from the ground to be future ready; its building blocks are constructed such that they handle enormous amounts of data. Ceph is a true unified storage solution that provides object, block, and file storage from a single unified software layer. When we call Ceph as future ready, we mean to focus on its object storage capabilities, which is a better fit for today's mix of unstructured data than blocks or files. Everything in Ceph relies on intelligent objects, whether it's block storage or file storage.
Introducing Ceph Storage

Rather than managing blocks and files underneath, Ceph manages objects and supports block- and file-based storage on top of it. If you think of a traditional file-based storage system, files are addressed via the file path, and in a similar way, objects in Ceph are addressed by a unique identifier, and are stored in a flat addressed space. Objects provide limitless scaling with increased performance by eliminating metadata operations. Ceph uses an algorithm to dynamically compute where the object should be stored and retrieved from.

The next generation architecture

The traditional storage systems do not have a smarter way of managing metadata. Metadata is the information (data) about data, which decides where the data will be written to and read from. Traditional storage systems maintain a central lookup table to keep track of their metadata; that is, every time a client sends a request for a read or write operation, the storage system first performs a lookup to the huge metadata table, and after receiving the results, it performs the client operation. For a smaller storage system, you might not notice performance hits, but think of a large storage cluster; you would definitely be restricted by performance limits with this approach. This would also restrict your scalability.

Ceph does not follow the traditional architecture of storage; it has been totally reinvented with the next-generation architecture. Rather than storing and manipulating metadata, Ceph introduces a newer way, the CRUSH algorithm. CRUSH stands for Controlled Replication Under Scalable Hashing. For more information, visit http://ceph.com/resources/publications/. Instead of performing a lookup in the metadata table for every client request, the CRUSH algorithm, on demand, computes where the data should be written to or read from. By computing metadata, there is no need to manage a centralized table for metadata. Modern computers are amazingly fast and can perform a CRUSH lookup very quickly; moreover, a smaller computing load can be distributed across cluster nodes, leveraging the power of distributed storage. CRUSH does clean management of metadata, which is a better way than the traditional storage system.

In addition to this, CRUSH has a unique property of infrastructure awareness. It understands the relationship between the various components of your infrastructure, right from the system disk, pool, node, rack, power board, switch, and data center row, to the data center room and further. These are failure zones for any infrastructure. CRUSH stores the primary copy of the data and its replica in a fashion such that data will be available even if a few components fail in a failure zone. Users have full control of defining these failure zones for their infrastructure inside Ceph's CRUSH map. This gives power to the Ceph administrator to efficiently manage the data of their own environment.
CRUSH makes Ceph self managing and self healing. In the event of component failure in a failure zone, CRUSH senses which component has failed and determines the effect of this failure on the cluster. Without any administrative intervention, CRUSH does self managing and self healing by performing a recovery operation for the data lost due to failure. CRUSH regenerates the data from the replica copies that the cluster maintains. At every point in time, the cluster will have more than one copy of data that will be distributed across the cluster.

Using CRUSH, we can design a highly reliable storage infrastructure with no single point of failure. It makes Ceph a highly scalable and reliable storage system, which is future ready.

**Raid – end of an era**

Raid technology has been the fundamental building block for storage systems for many years. It has proven successful for almost every kind of data that has been generated in the last 30 years. However, all eras must come to an end, and this time, it's for RAID. RAID-based storage systems have started to show limitations and are incapable of delivering future storage needs.

Disk-manufacturing technology is getting mature over the years. Manufacturers are now producing larger-capacity enterprise disks at lower prices. We no longer talk about 450 GB, 600 GB, or even 1 TB disks as there are a lot of other options with larger-capacity, better performing disks available today. The newer enterprise disk specifications offer up to 4 TB and even 6 TB disk drives. Storage capacity will keep on increasing year by year.

Think of an enterprise RAID-based storage system that is made up of numerous 4 or 6 TB disk drives; in the event of disk failure, RAID will take several hours and even up to days to repair a single failed disk. Meanwhile, if another drive fails, that would be chaos. Repairing multiple large disk drives using RAID is a cumbersome process.

Moreover, RAID eats up a lot of whole disks as a spare disk. This again affects the TCO, and if you are running short of spare disks, then again you are in trouble. The RAID mechanism requires a set of identical disks in a single RAID group; you will face penalties if you change the disk size, RPM, and disk type. Doing this will adversely affect the capacity and performance of your storage system.
Introducing Ceph Storage

Enterprise RAID-based systems often require expensive hardware component also known as RAID cards, which again increases the overall costs. RAID can hit a dead end when it’s not possible to grow its size that is no scale up or scale out feature after a certain limit. You cannot add more capacity even though you have the money. RAID 5 can survive a single disk failure and RAID 6 survives two-disk failure, which is the maximum for any RAID level. At the time of RAID recovery operations, if clients are performing an operation, they will most likely starve for I/O until the recovery operation finishes. The most limiting factor in RAID is that it only protects against disk failure; it cannot protect against failure of a network, server hardware, OS, switch, or regional disaster. The maximum protection you can get from RAID is survival for two-disk failures; you cannot survive more than two-disk failures in any circumstance.

Hence, we need a system that can overcome all these drawbacks in a performance-and cost-effective way. A Ceph storage system is the best solution available today to address these problems. For data reliability, Ceph makes use of the data replication method; that is, it does not use RAID, and because of this, it simply overcomes all the problems that can be found in a RAID-based enterprise system. Ceph is a software-defined storage, so we do not require any specialized hardware for data replication; moreover, the replication level is highly customized by means of commands; that is, the Ceph storage administrator can easily manage a replication factor as per their requirements and underlying infrastructure. In the event of one or more disk failures, Ceph's replication is a better process than that in RAID. When a disk drive fails, all the data that was residing on that disk at that point of time starts to recover from its peer disks. Since Ceph is a distributed system, all the primary copies and replicated copies of data are scattered on all the cluster disks such that no primary and replicated copy should reside on the same disk and must reside on a different failure zone defined by the CRUSH map. Hence, all the cluster disks participate in data recovery. This makes the recovery operation amazingly fast without performance bottlenecks. This recovery operation does not require any spare disk; data is simply replicated to other Ceph disks in the cluster. Ceph uses a weighting mechanism for its disks; hence, different disk sizes is not a problem. Ceph stores data based on the disk's weight, which is intelligently managed by Ceph and can also be managed by custom CRUSH maps.

In addition to the replication method, Ceph also supports another advance way of data reliability, by using the erasure-coding technique. Erasure-coded pools require less storage space compared to replicated pools. In this process, data is recovered or regenerated algorithmically by erasure-code calculation. You can use both the techniques of data availability, that is, replication as well as erasure coding, in the same Ceph cluster but over different storage pools. We will learn more about the erasure-coding technique in the coming chapters.
The compatibility portfolio

Ceph is an enterprise-ready storage system that offers support to a wide range of protocols and accessibility methods. The unified Ceph storage system supports block, file, and object storage; however, at the time of writing this book, Ceph block and object storage are recommended for production usage, and the Ceph filesystem is under QA testing and will be ready soon. We will discuss each of them in brief.

Ceph block storage

Block storage is a category of data storage used in the storage area network. In this type, data is stored as volumes, which are in the form of blocks and are attached to nodes. This provides a larger storage capacity required by applications with a higher degree of reliability and performance. These blocks, as volumes, are mapped to the operating system and are controlled by its filesystem layout.

Ceph has introduced a new protocol RBD that is now known as Ceph Block Device. RBD provides reliable, distributed, and high performance block storage disks to clients. RBD blocks are striped over numerous objects, which are internally scattered over the entire Ceph cluster, thus providing data reliability and performance to clients. RBD has native support for the Linux kernel. In other words, RBD drivers have been well integrated with the Linux kernel since the past few years. Almost all the Linux OS flavors have native support for RBD. In addition to reliability and performance, RBD also provides enterprise features such as full and incremental snapshots, thin provisioning, copy-on-write cloning, and several others. RBD also supports in-memory caching, which drastically improves its performance.

Ceph RBD supports images up to the size of 16 exabytes. These images can be mapped as disks to bare metal machines, virtual machines, or to a regular host machine. The industry-leading open source hypervisors such as KVM and Zen provide full support to RBD and leverage their features to their guest virtual machines. Other proprietary hypervisors such as VMware and Microsoft HyperV will be supported very soon. There has been a lot of work going on in the community for support to these hypervisors.
The Ceph block device provides full support to cloud platforms such as OpenStack, CloudStack, as well as others. It has been proven successful and feature-rich for these cloud platforms. In OpenStack, you can use the Ceph block device with the cinder (block) and glance (imaging) components; by doing this, you can spin 1,000s of VMs in very little time, taking advantage of the copy-on-write feature of the Ceph block storage.

The Ceph filesystem

The Ceph filesystem, also known as CephFS, is a POSIX-compliant filesystem that uses the Ceph storage cluster to store user data. CephFS has support for the native Linux kernel driver, which makes CephFS highly adaptive across any flavor of the Linux OS. CephFS stores data and metadata separately, thus providing increased performance and reliability to the application hosted on top of it.

Inside a Ceph cluster, the Ceph filesystem library (libcephfs) works on top of the RADOS library (librados), which is the Ceph storage cluster protocol, and is common for file, block, and object storage. To use CephFS, you will require at least one Ceph metadata server (MDS) to be configured on any of your cluster nodes. However, it's worth keeping in mind that only one MDS server will be a single point of failure for the Ceph filesystem. Once MDS is configured, clients can make use of CephFS in multiple ways. To mount Ceph as a filesystem, clients may use native Linux kernel capabilities or can make use of the ceph-fuse (filesystem in user space) drivers provided by the Ceph community.
In addition to this, clients can make use of third-party open source programs such as Ganesha for NFS and Samba for SMB/CIFS. These programs interact with libcephfs to store user's data to a reliable and distributed Ceph storage cluster. CephFS can also be used as a replacement for Apache Hadoop File System (HDFS). It also makes use of the libcephfs component to store data to the Ceph cluster. For its seamless implementation, the Ceph community provides the required CephFS Java interface for Hadoop and Hadoop plugins. The libcephfs and librados components are very flexible and you can even build your custom program that interacts with it and stores data to the underlying Ceph storage cluster.

CephFS is the only component of the Ceph storage system, which is not production-ready at the time of writing this book. It has been improving at a very high pace and is expected to be production-ready very soon. Currently, it's quite popular in the testing and development environment, and has been evolved with enterprise-demanding features such as dynamic rebalancing and a subdirectory snapshot. The following diagram shows various ways in which CephFS can be used:

**Ceph object storage**

Object storage is an approach to storing data in the form of objects rather than traditional files and blocks. Object-based storage has been getting a lot of industry attention. Organizations that look for flexibility for their enormous data are rapidly adopting object storage solutions. Ceph is known to be a true object-based storage system.

Ceph is a distributed object storage system, which provides an object storage interface via Ceph's object gateway, also known as the RADOS gateway (radosgw). The RADOS gateway uses libraries such as librgw (the RADOS gateway library) and librados, allowing applications to establish a connection with the Ceph object storage. Ceph delivers one of the most stable multitenant object storage solutions accessible via a RESTful API.
Introducing Ceph Storage

The RADOS gateway provides a RESTful interface to the user application to store data on the Ceph storage cluster. The RADOS gateway interfaces are:

- **Swift compatibility**: This is an object storage functionality for the OpenStack Swift API
- **S3 compatibility**: This is an object storage functionality for the Amazon S3 API
- **Admin API**: This is also known as the management API or native API, which can be used directly in the application to gain access to the storage system for management purposes

To access Ceph's object storage system, you can also bypass the RADOS gateway layer, thus making accessibility more flexible and quicker. The librados software libraries allow user applications to directly access Ceph object storage via C, C++, Java, Python, and PHP. Ceph object storage has multisite capabilities; that is, it provides solutions for disaster recovery. Multisite object storage configuration can be achieved by RADOS or by federated gateways. The following diagram shows different API systems that can be used with Ceph:

![Diagram of API systems with Ceph Cluster, RADOS Gateway, Direct Access, and RESTful HTTP/5 Access]

**Ceph versus others**

The storage market needs a shift; proprietary storage systems are incapable of providing future data storage needs at a relatively low budget. After hardware procurement, licensing, support, and management costs, the proprietary systems are very expensive. In contrast to this, the open source storage technologies are well proven for their performance, reliability, scalability, and lower TCO. Numerous organizations, government-owned as well as private, universities, research and healthcare centers, and HPC systems are already using some kind of open source storage solution.
However, Ceph is getting tremendous feedback and gaining popularity, leaving other open source as well as proprietary storage solutions behind. The following are some open source storage solutions in competition with Ceph. We will briefly discuss the shortcomings of these storage solutions, which have been addressed in Ceph.

**GPFS**

*General Parallel File System (GPFS)* is a distributed filesystem, developed and owned by IBM. This is a proprietary and closed source storage system, which makes it less attractive and difficult to adapt. The licensing and support cost after storage hardware makes it very expensive. Moreover, it has a very limited set of storage interfaces; it provides neither block storage nor RESTful access to the storage system, so this is very restrictive deal. Even the maximum data replication is limited to only three copies, which reduces system reliability in the event of more than one simultaneous failure.

**iRODS**

iRODS stands for *Integrated Rule-Oriented Data System*, which is an open source data-management software released under a 3-clause BSD license. iRods is not a highly reliable storage system as its iCAT metadata server is SPOF (single point of failure) and it does not provide true HA. Moreover, it has a very limited set of storage interfaces; it neither provides block storage nor RESTful access to the storage system, thus making it very restrictive. It's more suitable to store a small quantity of big files rather than both small and big files. iRods works in a traditional way, maintaining an index of the physical location, which is associated with the filename. The problem arises with multiple clients' request for the file location from the metadata server, applying more computing pressure on the metadata server, resulting in dependency on a single machine and performance bottlenecks.

**HDFS**

HDFS is a distributed scalable filesystem written in Java for the Hadoop framework. HDFS is not a fully POSIX-compliant filesystem and does not support block storage, thus making it less usable than Ceph. The reliability of HDFS is a question for discussion as it's not a highly available filesystem. The single NameNode in HDFS is the primary reason for its single point of failure and performance bottleneck problems. It's more suitable to store a small quantity of big files rather than both small and big files.
Introducing Ceph Storage

Lustre

Lustre is a parallel-distributed filesystem driven by the open source community and is available under GNU General Public License. In Lustre, a single server is responsible to store and manage metadata. Thus, the I/O request from the client is totally dependent on single server's computing power, which is quite low for an enterprise-level consumption. Like iRODS and HDFS, Lustre is suitable to store a small quantity of big files rather than both small and big files. Similar to iRODS, Lustre manages an index file that maintains physical addresses mapped with filenames, which makes its architecture traditional and prone to performance bottlenecks. Lustre does not have any mechanism for node failure detection and correction. In the event of node failure, clients have to connect to another node themselves.

Gluster

GlusterFS was originally developed by Gluster, which was then bought by Red Hat in 2011. GlusterFS is a scale-out network-attached filesystem. In Gluster, administrators have to determine which placement strategy to use to store data replica on different geographical racks. Gluster does not provide block access, filesystem, and remote replication as its intrinsic functions; rather, it provides these features as add-ons.

Ceph

If we make a comparison between Ceph and other storage solutions available today, Ceph clearly stands out of the crowd due to its feature set. It has been developed to overcome the limitations of existing storage systems, and it has proved to be an ideal replacement for old and expensive proprietary storage systems. It’s an open source, software-defined storage solution on top of any commodity hardware, which makes it an economic storage solution. Ceph provides a variety of interfaces for the clients to connect to a Ceph cluster, thus increasing flexibility for clients. For data protection, Ceph does not rely on RAID technology as it's getting limited due to various reasons mentioned earlier in this chapter. Rather, it uses replication and erasure coding, which have been proved to be better solutions than RAID.

Every component of Ceph is reliable and supports high availability. If you configure Ceph components by keeping redundancy in mind, we can confidently say that Ceph does not have any single point of failure, which is a major challenge for other storage solutions available today. One of the biggest advantages of Ceph is its unified nature, where it provides out-of-the-box block, file, and object storage solutions, while other storage systems are still incapable of providing such features. Ceph is suitable to store both small as well as big files without any performance glitch.
Ceph is a distributed storage system; clients can perform quick transactions using Ceph. It does not follow the traditional method of storing data, that is, maintaining metadata that is tied to a physical location and filename; rather, it introduces a new mechanism, which allows clients to dynamically calculate data location required by them. This gives a boost in performance for the client, as they no longer need to wait to get data locations and contents from the central metadata server. Moreover, the data placement inside the Ceph cluster is absolutely transparent and automatic; neither the client nor the administrators have to bother about data placement on a different failure zone. Ceph's intelligent system takes care of it.

Ceph is designed to self heal and self manage. In the event of disaster, when other storage systems cannot provide reliability against multiple failures, Ceph stands rock solid. Ceph detects and corrects failure at every failure zone such as a disk, node, network, rack, data center row, data center, and even different geographies. Ceph tries to manage the situation automatically and heal it wherever possible without data outage. Other storage solutions can only provide reliability up to disk or at node failure.

When it comes to a comparison, these are just a few features of Ceph to steal the show and stand out from the crowd.

**Summary**

Ceph is an open source software-defined storage solution that runs on commodity hardware, thus enabling enterprises to get rid of expensive, restrictive, proprietary storage systems. It provides a unified, distributed, highly scalable, and reliable object storage solution, which is much needed for today's and the future's unstructured data needs. The world's storage need is exploding, so we need a storage system that is scalable to the multiexabyte level without affecting data reliability and performance. Ceph is future proof, and provides a solution to all these data problems. Ceph is in demand for being a true cloud storage solution with support for almost every cloud platform. From every perspective, Ceph is a great storage solution available today.
Where to buy this book

You can buy Learning Ceph from the Packt Publishing website.

Alternatively, you can buy the book from Amazon, BN.com, Computer Manuals and most internet book retailers.

Click here for ordering and shipping details.