Ceph (pronounced /ˈsɛf/) is a free and open-source software-defined storage platform that provides object storage,[7] block storage, and file storage built on a common distributed cluster foundation. Ceph provides completely distributed operation without a single point of failure and scalability to the exabyte level, and is freely available. Since version 12 (Luminous), Ceph does not rely on any other conventional filesystem and directly manages HDDs and SSDs with its own storage backend BlueStore and can expose a POSIX filesystem.

Ceph Storage

Original author(s)	Inktank Storage (Sage Weil, Yehuda Sadeh Weinraub, Gregory Farnum, Josh Durgin, Samuel Just, Wido den Hollander)
Developer(s)	Red Hat, Intel, CERN, Cisco, Fujitsu, SanDisk, Canonical and SUSE[1]
Stable release	18.2.0[2] (Reef) / 3 August 2023
Repository	github.com/ceph/ceph
Written in	C++, Python[3]
Operating system	Linux, FreeBSD,[4] Windows[5]
Type	Distributed object store
License	LGPLv2.1[6]
Website	ceph.io

Ceph replicates data with fault tolerance,[8] using commodity hardware and Ethernet IP and requiring no specific hardware support. Ceph is highly available and ensures strong data durability through techniques including replication, erasure coding, snapshots and clones. By design, the system is both self-healing and self-managing, minimizing administration time and other costs.

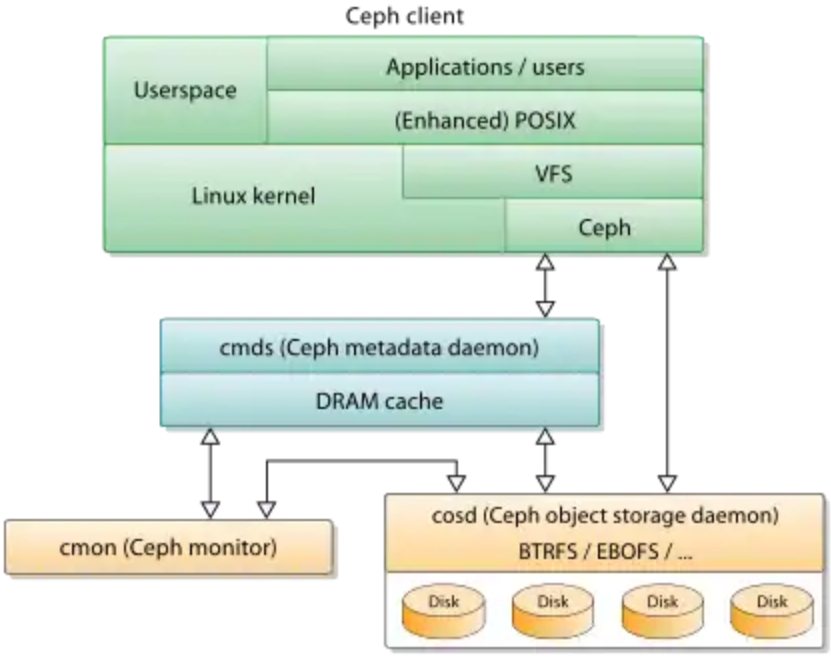
Large-scale production Ceph deployments include CERN,[9][10] OVH[11][12][13][14] and DigitalOcean.[15][16]

Design

Ceph employs five distinct kinds of daemons:[17]

- Cluster monitors (`ceph-mon`) that keep track of active and failed cluster nodes, cluster configuration, and information about data placement and global cluster state.
- OSDs (`ceph-osd`) that manage bulk data storage devices directly via the BlueStore back end,[18] which since the v12.x release replaces the Filestore[19] back end, which was implemented on top of a traditional filesystem)

- Metadata servers (`ceph-mds`) that maintain and broker access to inodes and directories inside a CephFS filesystem
- HTTP gateways (`ceph-rgw`) that expose the object storage layer as an interface compatible with Amazon S3 or OpenStack Swift APIs
- Managers (`ceph-mgr`) that perform cluster monitoring, bookkeeping, and maintenance tasks, and interface to external monitoring systems and management (e.g. balancer, dashboard, Prometheus, Zabbix plugin)[20]



A high-level overview of the Ceph's internal organization[17]:4

All of these are fully distributed, and may be deployed on disjoint, dedicated servers or in a converged topology. Clients with different needs directly interact with appropriate cluster components.[21]

Ceph distributes data across multiple storage devices and nodes to achieve higher throughput, in a fashion similar to RAID. Adaptive load balancing is supported whereby frequently accessed services may be replicated over more nodes.[22]

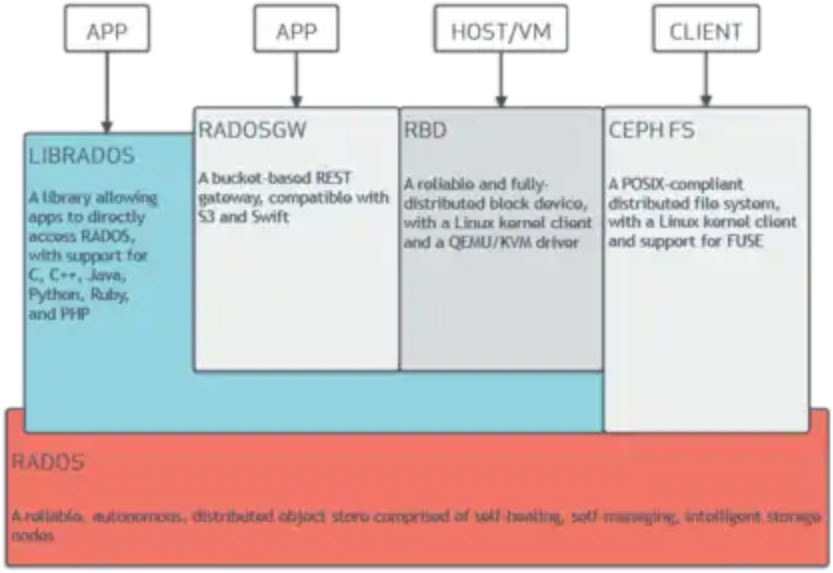
As of September 2017, BlueStore is the default and recommended storage back end for production environments,[23] which provides better latency and configurability than the older Filestore back end, and avoiding the shortcomings of filesystem based storage involving additional processing and caching layers. The Filestore back end will be deprecated as of the Reef release in mid 2023. XFS was the recommended underlying filesystem for Filestore OSDs, and Btrfs could be used at one's own risk. ext4 filesystems were not recommended due to limited metadata capacity.[24] The BlueStore back end does still use XFS for a small metadata partition.[25]

Object storage S3

Ceph implements distributed object storage via the RADOS GateWay (`ceph-rgw`), which exposes the underlying storage layer via an interface compatible with Amazon S3 or

OpenStack Swift.

Ceph RGW deployments scale readily and often utilize large and dense storage media for bulk use cases that include Big Data (datalake), backups & archives, IOT, media, video recording, and deployment images for virtual machines and containers.[26]



An architecture diagram showing the relations among components of the Ceph storage platform

Ceph's software libraries provide client applications with direct access to the *reliable autonomic distributed object store* (RADOS) object-based storage system. More frequently used are libraries for Ceph's *RADOS Block Device* (RBD), *RADOS Gateway*, and *Ceph File System* services. In this way, administrators can maintain their storage devices within a unified system, which makes it easier to replicate and protect the data.

The "librados" software libraries provide access in C, C++, Java, PHP, and Python. The RADOS Gateway also exposes the object store as a RESTful interface which can present as both native Amazon S3 and OpenStack Swift APIs.

Block storage

Ceph can provide clients with thin-provisioned block devices. When an application writes data to Ceph using a block device, Ceph automatically stripes and replicates the data across the cluster. Ceph's *RADOS Block Device* (RBD) also integrates with Kernel-based Virtual Machines (KVMs).

Ceph block storage may be deployed on traditional HDDs and/or SSDs which are associated with Ceph's block storage for use cases, including databases, virtual machines, data analytics, artificial intelligence, and machine learning. Block storage clients often require high throughput and IOPS, thus Ceph RBD deployments increasingly utilize SSDs with NVMe interfaces.

"RBD" is built on with Ceph's foundational RADOS object storage system that provides the librados interface and the CephFS file system. Since RBD is built on librados, RBD

inherits librados's abilities, including clones and snapshots. By striping volumes across the cluster, Ceph improves performance for large block device images.

"Ceph-iSCSI" is a gateway which enables access to distributed, highly available block storage from Microsoft Windows and VMware vSphere servers or clients capable of speaking the iSCSI protocol. By using ceph-iscsi on one or more iSCSI gateway hosts, Ceph RBD images become available as Logical Units (LUs) associated with iSCSI targets, which can be accessed in an optionally load-balanced, highly available fashion.

Since ceph-iscsi configuration is stored in the Ceph RADOS object store, ceph-iscsi gateway hosts are inherently without persistent state and thus can be replaced, augmented, or reduced at will. As a result, Ceph Storage enables customers to run a truly distributed, highly-available, resilient, and self-healing enterprise storage technology on commodity hardware and an entirely open source platform.

The block device can be virtualized, providing block storage to virtual machines, in virtualization platforms such as Openshift, OpenStack, Kubernetes, OpenNebula, Ganeti, Apache CloudStack and Proxmox Virtual Environment.

File storage

Ceph's file system (CephFS) runs on top of the same RADOS foundation as Ceph's object storage and block device services. The CephFS metadata server (MDS) provides a service that maps the directories and file names of the file system to objects stored within RADOS clusters. The metadata server cluster can expand or contract, and it can rebalance file system metadata ranks dynamically to distribute data evenly among cluster hosts. This ensures high performance and prevents heavy loads on specific hosts within the cluster.

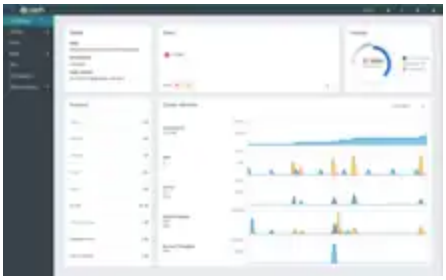
Clients mount the POSIX-compatible file system using a Linux kernel client. An older FUSE-based client is also available. The servers run as regular Unix daemons.

Ceph's file storage is often associated with log collection, messaging, and file storage.

Dashboard

From 2018 there is also a Dashboard web UI project, which helps to manage the cluster. It's being developed by Ceph community on LGPL-3 and uses Ceph-mgr, Python, Angular framework and Grafana.[27] Landing page has been refreshed in the beginning of 2023.[28]

Previous dashboards were developed but are closed now: Calamari (2013-2018), OpenAttic (2013-2019), VSM (2014-2016), Inkscope (2015-2016) and Ceph-Dash (2015-2017).[29]



Ceph Dashboard landing page (2023)

Crimson

Beginning in 2019 the Crimson project has been reimplementing the OSD data path. The goal of Crimson is to minimize latency and CPU overhead. Modern storage devices and interfaces including NVMe and 3D_XPoint have become much faster than HDD and even SAS/SATA SSDs, but CPU performance has not kept pace. Moreover `crimson-osd` is meant to be a backward-compatible drop-in replacement for `ceph-osd`. While Crimson can work with the BlueStore back end (via AlienStore), a new native ObjectStore implementation called SeaStore is also being developed along with CyanStore for testing purposes. One reason for creating SeaStore is that transaction support in the BlueStore back end is provided by RocksDB, which needs to be re-implemented to achieve better parallelism.[30][31][32]

History

Ceph was created by Sage Weil for his doctoral dissertation,[33] which was advised by Professor Scott A. Brandt at the Jack Baskin School of Engineering, University of California, Santa Cruz (UCSC), and sponsored by the Advanced Simulation and Computing Program (ASC), including Los Alamos National Laboratory (LANL), Sandia National Laboratories (SNL), and Lawrence Livermore National Laboratory (LLNL).[34] The first line of code that ended up being part of Ceph was written by Sage Weil in 2004 while at a summer internship at LLNL, working on scalable filesystem metadata management (known today as Ceph's MDS).[35] In 2005, as part of a summer project initiated by Scott A. Brandt and led by Carlos Maltzahn, Sage Weil created a fully functional file system prototype which adopted the name Ceph. Ceph made its debut with Sage Weil giving two presentations in November 2006, one at USENIX OSDI 2006[36] and another at SC'06.[37]

After his graduation in autumn 2007, Weil continued to work on Ceph full-time, and the core development team expanded to include Yehuda Sadeh Weinraub and Gregory Farnum. On March 19, 2010, Linus Torvalds merged the Ceph client into Linux kernel

version 2.6.34[38][39] which was released on May 16, 2010. In 2012, Weil created Inktank Storage for professional services and support for Ceph.[40][41]

In April 2014, Red Hat purchased Inktank, bringing the majority of Ceph development in-house to make it a production version for enterprises with support (hotline) and continuous maintenance (new versions).[42]

In October 2015, the Ceph Community Advisory Board was formed to assist the community in driving the direction of open source software-defined storage technology. The charter advisory board includes Ceph community members from global IT organizations that are committed to the Ceph project, including individuals from Red Hat, Intel, Canonical, CERN, Cisco, Fujitsu, SanDisk, and SUSE.[43]

In November 2018, the Linux Foundation launched the Ceph Foundation as a successor to the Ceph Community Advisory Board. Founding members of the Ceph Foundation included Amihan, Canonical, China Mobile, DigitalOcean, Intel, OVH, ProphetStor Data Services, Red Hat, SoftIron, SUSE, Western Digital, XSKY Data Technology, and ZTE.[44]

In March 2021, SUSE discontinued its Enterprise Storage product incorporating Ceph in favor of Longhorn,[45] and the former Enterprise Storage website was updated stating "SUSE has refocused the storage efforts around serving our strategic SUSE Enterprise Storage Customers and are no longer actively selling SUSE Enterprise Storage."[46]

Release history

Release history

Name	Release	First release	End of life	Milestones
Argonaut	0.48	July 3, 2012		First major "stable" release
Bobtail	0.56	January 1, 2013		
Cuttlefish	0.61	May 7, 2013		ceph-deploy is stable
Dumpling	0.67	August 14, 2013	May 2015	namespace, region, monitoring REST API
Emperor	0.72	November 9, 2013	May 2014	multi-datacenter replication for RGW
Firefly	0.80	May 7, 2014	April 2016	erasure coding, cache tiering, primary affinity, key/value OSD backend (experimental), standalone RGW (experimental)
Giant	0.87	October 29, 2014	April 2015	
Hammer	0.94	April 7, 2015	August 2017	
Infernalis	9.2.0	November 6, 2015	April 2016	
Jewel	10.2.0	April 21, 2016	2018-06-01	Stable CephFS, experimental OSD back end named BlueStore, daemons no longer run as the root user
Kraken	11.2.0	January 20, 2017	2017-08-01	BlueStore is stable, EC for RBD pools
Luminous	12.2.0	August 29, 2017	2020-03-01	pg-upmap balancer
Mimic	13.2.0	June 1, 2018	2020-07-22	snapshots are stable, Beast is stable, official GUI (Dashboard)

Nautilus	14.2.0	March 19, 2019	2021-06-01	asynchronous replication, auto-retry of failed writes due to grown defect remapping
Octopus	15.2.0	March 23, 2020	2022-06-01	
Pacific	16.2.0	March 31, 2021[47]	2023-06-01	
Quincy	17.2.0	April 19, 2022[48]	2024-06-01	auto-setting of min_alloc_size for novel media
Reef	**18.2.0**	Aug 3, 2023[49]		
Squid	TBA	TBA		

Legend: Old version Older version, still maintained **Latest version** Future release

Available platforms

While basically built for Linux, Ceph has been also partially ported to Windows platform. It is production-ready for Windows Server 2016 (some commands might be unavailable due to lack of UNIX socket implementation), Windows Server 2019 and Windows Server 2022, but testing/development can be done also on Windows 10 and Windows 11. One can use Ceph RBD and CephFS on Windows, but OSD is not supported on this platform.[50][5][51]

There is also FreeBSD implementation of Ceph.[4]

Etymology

The name "Ceph" is a shortened form of "cephalopod", a class of molluscs that includes squids, cuttlefish, nautiloids, and octopuses. The name (emphasized by the logo) suggests the highly parallel behavior of an octopus and was chosen to associate the file system with "Sammy", the banana slug mascot of UCSC.[17] Both cephalopods and banana slugs are molluscs.

See also

- BeeGFS
- Distributed file system
- Distributed parallel fault-tolerant file systems
- Gfarm file system
- GlusterFS
- IBM General Parallel File System (GPFS)
- Kubernetes
- LizardFS
- Lustre
- MapR FS
- Moose File System
- OrangeFS
- Parallel Virtual File System
- Quantcast File System
- RozoFS
- Software-defined storage
- XtreemFS
- ZFS
- Comparison of distributed file systems

References

1. ↑ "Ceph Community Forms Advisory Board". 2015-10-28. Archived from the original on 2019-01-29. Retrieved 2016-01-20.

2. ↑ Error: Unable to display the reference properly. See the documentation for details.

3. ↑ "GitHub Repository". *GitHub*.

4. 1 2 "FreeBSD Quarterly Status Report".

5. 1 2 "Installing Ceph on Windows". *Ceph*. Retrieved 2 July 2023.

6. ↑ "LGPL2.1 license file in the Ceph sources". *GitHub*. 2014-10-24. Retrieved 2014-10-24.

7. ↑ Nicolas, Philippe (2016-07-15). "The History Boys: Object storage ... from the beginning". *The Register*.

8. ↑ Jeremy Andrews (2007-11-15). "Ceph Distributed Network File System". KernelTrap. Archived from the original on 2007-11-17. Retrieved 2007-11-15.

9. ↑ "Ceph Clusters". *CERN*. Retrieved 12 November 2022.

10. ↑ "Ceph Operations at CERN: Where Do We Go From Here? - Dan van der Ster & Teo Mouratidis, CERN". *YouTube*. 24 May 2019. Retrieved 12 November 2022.

11. ↑ Dorosz, Filip (15 June 2020). "Journey to next-gen Ceph storage at OVHcloud with LXD". *OVHcloud*. Retrieved 12 November 2022.

12. ↑ "CephFS distributed filesystem". *OVHcloud*. Retrieved 12 November 2022.

13. ↑ "Ceph - Distributed Storage System in OVH [en] - Bartłomiej Święcki". *YouTube*. 7 April 2016. Retrieved 12 November 2022.

14. ↑ "200 Clusters vs 1 Admin - Bartosz Rabiega, OVH". *YouTube*. 24 May 2019. Retrieved 15 November 2022.

15. ↑ D'Atri, Anthony (31 May 2018). "Why We Chose Ceph to Build Block Storage". *DigitalOcean*. Retrieved 12 November 2022.

16. ↑ "Ceph Tech Talk: Ceph at DigitalOcean". *YouTube*. 7 October 2021. Retrieved 12 November 2022.

17. 1 2 3 M. Tim Jones (2010-06-04). "Ceph: A Linux petabyte-scale distributed file system" (PDF). IBM. Retrieved 2014-12-03.

18. ↑ "BlueStore". Ceph. Retrieved 2017-09-29.

19. ↑ "BlueStore Migration". Archived from the original on 2019-12-04. Retrieved 2020-04-12.

20. ↑ "Ceph Manager Daemon — Ceph Documentation". *docs.ceph.com*. Archived from the original on June 6, 2018. Retrieved 2019-01-31. archive link Archived June 19, 2020, at the Wayback Machine

21. ↑ Jake Edge (2007-11-14). "The Ceph filesystem". LWN.net.

22. ↑ Anthony D'Atri, Vaibhav Bhembre (2017-10-01). "Learning Ceph, Second Edition". Packt.

23. ↑ Sage Weil (2017-08-29). "v12.2.0 Luminous Released". Ceph Blog.

24. ↑ "Hard Disk and File System Recommendations". ceph.com. Archived from the original on 2017-07-14. Retrieved 2017-06-26.

25. ↑ "BlueStore Config Reference". Archived from the original on July 20, 2019. Retrieved April 12, 2020.

26. ↑ "10th International Conference "Distributed Computing and Grid Technologies in Science and Education" (GRID'2023)". *JINR (Indico)*. 2023-07-03. Retrieved 2023-08-09.

27. ↑ "Ceph Dashboard". *Ceph documentation*. Retrieved 11 April 2023.

28. ↑ Gomez, Pedro Gonzalez (23 February 2023). "Introducing the new Dashboard Landing Page". Retrieved 11 April 2023.

29. ↑ "Operating Ceph from the Ceph Dashboard: past, present and future". *YouTube*. 22 November 2022. Retrieved 11 April 2023.

30. ↑ Just, Sam (18 January 2021). "Crimson: evolving Ceph for high performance NVMe". *Red Hat Emerging Technologies*. Retrieved 12 November 2022.

31. ↑ Just, Samuel (10 November 2022). "What's new with Crimson and Seastore?". *YouTube*. Retrieved 12 November 2022.

32. ↑ "Crimson: Next-generation Ceph OSD for Multi-core Scalability". *Ceph blog*. Ceph. 7 February 2023. Retrieved 11 April 2023.

33. ↑ Sage Weil (2007-12-01). "Ceph: Reliable, Scalable, and High-Performance Distributed Storage" (PDF). University of California, Santa Cruz. Archived from the original (PDF) on 2017-07-06. Retrieved 2017-03-11.

34. ↑ Gary Grider (2004-05-01). "The ASCI/DOD Scalable I/O History and Strategy" (PDF). *University of Minnesota*. Retrieved 2019-07-17.

35. ↑ Dynamic Metadata Management for Petabyte-Scale File Systems, SA Weil, KT Pollack, SA Brandt, EL Miller, Proc. SC'04, Pittsburgh, PA, November, 2004

36. ↑ "Ceph: A scalable, high-performance distributed file system," SA Weil, SA Brandt, EL Miller, DDE Long, C Maltzahn, Proc. OSDI, Seattle, WA, November, 2006

37. ↑ "CRUSH: Controlled, scalable, decentralized placement of replicated data," SA Weil, SA Brandt, EL Miller, DDE Long, C Maltzahn, SC'06, Tampa, FL, November, 2006

38. ↑ Sage Weil (2010-02-19). "Client merged for 2.6.34". ceph.newdream.net. Archived from the original on 2010-03-23. Retrieved 2010-03-21.

39. ↑ Tim Stephens (2010-05-20). "New version of Linux OS includes Ceph file system developed at UCSC". news.ucsc.edu.

40. ↑ Bryan Bogensberger (2012-05-03). "And It All Comes Together". Inktank Blog. Archived from the original on 2012-07-19. Retrieved 2012-07-10.

41. ↑ Joseph F. Kovar (July 10, 2012). "The 10 Coolest Storage Startups Of 2012 (So Far)". *CRN*. Retrieved July 19, 2013.

42. ↑ Red Hat Inc (2014-04-30). "Red Hat to Acquire Inktank, Provider of Ceph". Red Hat. Retrieved 2014-08-19.

43. ↑ "Ceph Community Forms Advisory Board". 2015-10-28. Archived from the original on 2019-01-29. Retrieved 2016-01-20.

44. ↑ "The Linux Foundation Launches Ceph Foundation To Advance Open Source Storage". 2018-11-12.

45. ↑ "SUSE says tschüss to Ceph-based enterprise storage product – it's Rancher's Longhorn from here on out".

46. ↑ "SUSE Enterprise Software-Defined Storage".

47. ↑ Ceph.io — v16.2.0 Pacific released

48. ↑ Ceph.io — v17.2.0 Quincy released

49. ↑ Flores, Laura (6 August 2023). "v18.2.0 Reef released". *Ceph Blog*. Retrieved 26 August 2023.

50. ↑ "Ceph for Windows". Cloudbase Solutions. Retrieved 2 July 2023.

51. ↑ Pilotti, Alessandro. "Ceph on Windows". *YouTube*. Retrieved 2 July 2023.

Further reading

- M. Tim Jones (2010-05-04). "Ceph: A Linux petabyte-scale distributed file system". *DeveloperWorks > Linux > Technical Library*. Retrieved 2010-05-06.
- Jeffrey B. Layton (2010-04-20). "Ceph: The Distributed File System Creature from the Object Lagoon". *Linux Magazine*. Archived from the original on April 23, 2010. Retrieved 2010-04-24.
- Carlos Maltzahn; Esteban Molina-Estolano; Amandeep Khurana; Alex J. Nelson; Scott A. Brandt; Sage Weil (August 2010). "Ceph as a scalable alternative to the Hadoop Distributed File System". *;login:*. **35** (4). Retrieved 2012-03-09.
- Martin Loschwitz (April 24, 2012). "The RADOS Object Store and Ceph Filesystem". *HPC ADMIN Magazine*. Retrieved 2012-04-25.

External links

- Official website
- State of the Cephalopod 2022 on YouTube