V0.66

UPGRADING

- There is now a configurable maximum rados object size, defaulting to 100 GB. If you are using librados and storing objects larger than that, you will need to adjust 'osd max object size', and should consider using smaller objects instead.

NOTABLE CHANGES

- osd: pg log (re)writes are not vastly more efficient (faster peering) (Sam Just)
- osd: fixed problem with front-side heartbeats and mixed clusters (David Zafman)
- mon: tuning, performance improvements
- mon: simplify PaxosService vs Paxos interaction, fix readable/writeable checks
- rgw: fix radosgw-admin buckets list (Yehuda Sadeh)
- mds: support robust lookup by ino number (good for NFS) (Yan, Zheng)
- mds: fix several bugs (Yan, Zheng)
- ceph-fuse, libcephfs: fix truncatation bug on >4MB files (Yan, Zheng)
- ceph/librados: fix resending of commands on mon reconnect
- librados python: fix xattrs > 4KB (Josh Durgin)
- librados: configurable max object size (default 100 GB)
- msgr: fix various memory leaks
- ceph-fuse: fixed long-standing O_NOATIME vs O_LAZY bug
- ceph-fuse, libcephfs: fix request refcounting bug (hang on shutdown)
- ceph-fuse, libcephfs: fix read zeroing at EOF
- ceph-conf: –show-config-value now reflects daemon defaults
- ceph-disk: simpler, more robust locking
- ceph-disk: avoid mounting over an existing osd in /var/lib/ceph/osd/*
- sysvinit: handle symlinks in /var/lib/ceph/osd/*

V0.65

UPGRADING

- Huge revamp of the 'ceph' command-line interface implementation. The `ceph-common` client library needs to be upgrade before `ceph-mon` is restarted in order to avoid problems using the CLI (the old `ceph` client utility cannot talk to the new `ceph-mon`).

- The CLI is now very careful about sending the 'status' one-liner output to stderr and command output to stdout. Scripts relying on output should take care.

- The 'ceph osd tell ...' and 'ceph mon tell ...' commands are no longer supported. Any callers should use:

```
ceph tell osd.<id or *> ...
ceph tell mon.<id or name or *> ...
```

  The 'ceph mds tell ...' command is still there, but will soon also transition to 'ceph tell mds.<id or name or *> ...'

- The 'ceph osd crush add ...' command used to take one of two forms:

```
ceph osd crush add 123 osd.123 <weight> <location ...>
ceph osd crush add osd.123 <weight> <location ...>
```

  This is because the id and crush name are redundant. Now only the simple form is supported, where the osd name/id can either be a bare id (integer) or name (osd.<id>):

```
ceph osd crush add osd.123 <weight> <location ...>
ceph osd crush add 123 <weight> <location ...>
```

- There is now a maximum RADOS object size, configurable via ‘osd max object size’, defaulting to 100 GB. Note that this has no effect on RBD, CephFS, or radosgw, which all stripe over objects.

NOTABLE CHANGES

- mon, ceph: huge revamp of CLI and internal admin API. (Dan Mick)
- mon: new capability syntax
- osd: do not use fadvise(DONTNEED) on XFS (data corruption on power cycle)
- osd: recovery and peering performance improvements
- osd: new writeback throttling (for less bursty write performance) (Sam Just)
- osd: ping/heartbeat on public and private interfaces
- osd: avoid osd flapping from asymmetric network failure
- osd: re-use partially deleted PG contents when present (Sam Just)
- osd: break blacklisted client watches (David Zafman)
- mon: many stability fixes (Joao Luis)
- mon, osd: many memory leaks fixed
- mds: misc stability fixes (Yan, Zheng, Greg Farnum)
- mds: many backpointer improvements (Yan, Zheng)
- mds: new robust open-by-ino support (Yan, Zheng)
- ceph-fuse, libcephfs: fix a few caps revocation bugs
- librados: new calls to administer the cluster
- librbd: locking tests (Josh Durgin)
- ceph-disk: improved handling of odd device names
- ceph-disk: many fixes for RHEL/CentOS, Fedora, wheezy
- many many fixes from static code analysis (Danny Al-Gaaf)
- daemons: create /var/run/ceph as needed

V0.64

UPGRADING

- New pools now have the HASHPSPOOL flag set by default to provide better distribution over OSDs. Support for this feature was introduced in v0.59 and Linux kernel version v3.9. If you wish to access the cluster from an older kernel, set the ‘osd pool default flag hashpspool = false’ option in your ceph.conf prior to creating the cluster or creating new pools. Note that the presense of any pool in the cluster with the flag enabled will make the OSD require support from all clients.

NOTABLE CHANGES

- osd: monitor both front and back interfaces
- osd: verify both front and back network are working before rejoining cluster
- osd: fix memory/network inefficiency during deep scrub
- osd: fix incorrect mark-down of osds
- mon: fix start fork behavior
- mon: fix election timeout
- mon: better trim/compaction behavior
- mon: fix units in ‘ceph df’ output
- mon, osd: misc memory leaks
- librbd: make default options/features for newly created images (e.g., via qemu-img) configurable
- mds: many fixes for mds clustering
- mds: fix rare hang after client restart
- ceph-fuse: add ioctl support
- ceph-fuse/libcephfs: fix for cap release/hang
- rgw: handle deep uri resources
- rgw: fix CORS bugs
- ceph-disk: add ‘[un]suppress-active DEV’ command
- debian: rgw: stop daemon on uninstall
- debian: fix upstart behavior with upgrades

V0.63

UPGRADING

- The ‘osd min down {reporters|reports}’ config options have been renamed to ‘mon osd min down {reporters|reports}’, and the documentation has been updated to reflect that these options apply to the monitors (who process failure reports) and not OSDs. If you have adjusted these settings, please update your ``ceph.conf’’ accordingly.

NOTABLE CHANGES

- librbd: parallelize delete, rollback, flatten, copy, resize
- librbd: ability to read from local replicas
- osd: resurrect partially deleted PGs
- osd: prioritize recovery for degraded PGs
- osd: fix internal heartbeart timeouts when scrubbing very large objects
- osd: close narrow journal race
- rgw: fix usage log scanning for large, untrimmed logs
- rgw: fix locking issue, user operation mask,
- initscript: fix osd crush weight calculation when using -a
- initscript: fix enumeration of local daemons
- mon: several fixes to paxos, sync
- mon: new –extract-monmap to aid disaster recovery
- mon: fix leveldb compression, trimming
- add ‘config get’ admin socket command
- rados: clonedata command for cli
- debian: stop daemons on uninstall; fix dependencies
- debian wheezy: fix udev rules
- many many small fixes from coverity scan

V0.62

NOTABLE CHANGES

- mon: fix validation of mds ids from CLI commands
- osd: fix for an op ordering bug
- osd, mon: optionally dump leveldb transactions to a log
- osd: fix handling for split after upgrade from bobtail
- debian, specfile: packaging cleanups
- radosgw-admin: create keys for new users by default
- librados python binding cleanups
- misc code cleanups

V0.61.4 “CUTTLEFISH”

This release resolves a possible data corruption on power-cycle when using XFS, a few outstanding problems with monitor sync, several problems with ceph-disk and ceph-deploy operation, and a problem with OSD memory usage during scrub.

UPGRADING

- No issues.

NOTABLE CHANGES

- mon: fix daemon exit behavior when error is encountered on startup
- mon: more robust sync behavior
- osd: do not use sync_file_range(2), posix_fadvise(...DONTNEED) (can cause data corruption on power loss on XFS)
- osd: avoid unnecessary log rewrite (improves peering speed)

- osd: fix scrub efficiency bug (problematic on old clusters)
- rgw: fix listing objects that start with underscore
- rgw: fix deep URI resource, CORS bugs
- librados python binding: fix truncate on 32-bit architectures
- ceph-disk: fix udev rules
- rpm: install sysvinit script on package install
- ceph-disk: fix OSD start on machine reboot on Debian wheezy
- ceph-disk: activate OSD when journal device appears second
- ceph-disk: fix various bugs on RHEL/CentOS 6.3
- ceph-disk: add ‘zap’ command
- ceph-disk: add ‘[un]suppress-activate’ command for preparing spare disks
- upstart: start on runlevel [2345] (instead of after the first network interface starts)
- ceph-fuse, libcephfs: handle mds session reset during session open
- ceph-fuse, libcephfs: fix two capability revocation bugs
- ceph-fuse: fix thread creation on startup
- all daemons: create /var/run/ceph directory on startup if missing

For more detailed information, see **the complete changelog**.

V0.61.3 “CUTTLEFISH”

This release resolves a number of problems with the monitors and leveldb that users have been seeing. Please upgrade.

UPGRADING

- There is one known problem with mon upgrades from bobtail. If the ceph-mon conversion on startup is aborted or fails for some reason, we do not correctly error out, but instead continue with (in certain cases) odd results. Please be careful if you have to restart the mons during the upgrade. A 0.61.4 release with a fix will be out shortly.
- In the meantime, for current cuttlefish users, v0.61.3 is safe to use.

NOTABLE CHANGES

- mon: paxos state trimming fix (resolves runaway disk usage)
- mon: finer-grained compaction on trim
- mon: discard messages from disconnected clients (lowers load)
- mon: leveldb compaction and other stats available via admin socket
- mon: async compaction (lower overhead)
- mon: fix bug incorrectly marking osds down with insufficient failure reports
- osd: fixed small bug in pg request map
- osd: avoid rewriting pg info on every osdmap
- osd: avoid internal heartbeta timeouts when scrubbing very large objects
- osd: fix narrow race with journal replay
- mon: fixed narrow pg split race
- rgw: fix leaked space when copying object
- rgw: fix iteration over large/untrimmed usage logs
- rgw: fix locking issue with ops log socket
- rgw: require matching version of librados
- librbd: make image creation defaults configurable (e.g., create format 2 images via qemu-img)
- fix units in ‘ceph df’ output
- debian: fix prerm/postinst hooks to start/stop daemons appropriately
- upstart: allow uppercase daemons names (and thus hostnames)
- sysvinit: fix enumeration of local daemons by type
- sysvinit: fix osd weight calcuation when using -a
- fix build on unsigned char platforms (e.g., arm)

For more detailed information, see **the complete changelog**.

V0.61.2 “CUTTLEFISH”

This release disables a monitor debug log that consumes disk space and fixes a bug when upgrade some monitors from bobtail to cuttlefish.

NOTABLE CHANGES

- mon: fix conversion of stores with duplicated GV values
- mon: disable 'mon debug dump transactions' by default

For more detailed information, see **the complete changelog**.

V0.61.1 "CUTTLEFISH"

This release fixes a problem when upgrading a bobtail cluster that had snapshots to cuttlefish.

NOTABLE CHANGES

- osd: handle upgrade when legacy snap collections are present; repair from previous failed restart
- ceph-create-keys: fix race with ceph-mon startup (which broke 'ceph-deploy gatherkeys ...')
- ceph-create-keys: gracefully handle bad response from ceph-osd
- sysvinit: do not assume default osd_data when automatically weighting OSD
- osd: avoid crash from ill-behaved classes using getomapvals
- debian: fix squeeze dependency
- mon: debug options to log or dump leveldb transactions

For more detailed information, see **the complete changelog**.

V0.61 "CUTTLEFISH"

UPGRADING FROM V0.60

- The ceph-deploy tool is now the preferred method of provisioning new clusters. For existing clusters created via mkcephfs that would like to transition to the new tool, there is a migration path, documented at Transitioning to ceph-deploy.
- The sysvinit script (/etc/init.d/ceph) will now verify (and, if necessary, update) the OSD's position in the CRUSH map on startup. (The upstart script has always worked this way.) By default, this ensures that the OSD is under a 'host' with a name that matches the hostname (`hostname -s`). Legacy clusters create with mkcephfs do this by default, so this should not cause any problems, but legacy clusters with customized CRUSH maps with an alternate structure should set `osd crush update on start = false`.
- radosgw-admin now uses the term zone instead of cluster to describe each instance of the radosgw data store (and corresponding collection of radosgw daemons). The usage for the radosgw-admin command and the 'rgw zone root pool' config options have changed accordingly.
- rbd progress indicators now go to standard error instead of standard out. (You can disable progress with –no-progress.)
- The 'rbd resize ...' command now requires the –allow-shrink option when resizing to a smaller size. Expanding images to a larger size is unchanged.
- Please review the changes going back to 0.56.4 if you are upgrading all the way from bobtail.
- The old 'ceph stop_cluster' command has been removed.
- The sysvinit script now uses the ceph.conf file on the remote host when starting remote daemons via the '-a' option. Note that if '-a' is used in conjunction with '-c path', the path must also be present on the remote host (it is not copied to a temporary file, as it was previously).

UPGRADING FROM V0.56.4 "BOBTAIL"

Please see Upgrading from Bobtail to Cuttlefish for details.

- The ceph-deploy tool is now the preferred method of provisioning new clusters. For existing clusters created via mkcephfs that would like to transition to the new tool, there is a migration path, documented at Transitioning to ceph-deploy.

- The sysvinit script (/etc/init.d/ceph) will now verify (and, if necessary, update) the OSD's position in the CRUSH map on startup. (The upstart script has always worked this way.) By default, this ensures that the OSD is under a 'host' with a name that matches the hostname (`hostname -s`). Legacy clusters create with mkcephfs do this by default, so this should not cause any problems, but legacy clusters with customized CRUSH maps with an alternate structure should set `osd crush update on start = false`.
- radosgw-admin now uses the term zone instead of cluster to describe each instance of the radosgw data store (and

corresponding collection of radosgw daemons). The usage for the radosgw-admin command and the 'rgw zone root pool' config optoins have changed accordingly.
- rbd progress indicators now go to standard error instead of standard out. (You can disable progress with –no-progress.)
- The 'rbd resize ...' command now requires the –allow-shrink option when resizing to a smaller size. Expanding images to a larger size is unchanged.
- Please review the changes going back to 0.56.4 if you are upgrading all the way from bobtail.
- The old 'ceph stop_cluster' command has been removed.
- The sysvinit script now uses the ceph.conf file on the remote host when starting remote daemons via the '-a' option. Note that if '-a' is used in conjuction with '-c path', the path must also be present on the remote host (it is not copied to a temporary file, as it was previously).
- The monitor is using a completely new storage strategy and intra-cluster protocol. This means that cuttlefish and bobtail monitors do not talk to each other. When you upgrade each one, it will convert its local data store to the new format. Once you upgrade a majority, the quorum will be formed using the new protocol and the old monitors will be blocked out until they too get upgraded. For this reason, we recommend not running a mixed-version cluster for very long.
- ceph-mon now requires the creation of its data directory prior to –mkfs, similarly to what happens on ceph-osd. This directory is no longer automatically created, and custom scripts should be adjusted to reflect just that.
- The monitor now enforces that MDS names be unique. If you have multiple daemons start with with the same id (e.g., `mds.a`) the second one will implicitly mark the first as failed. This makes things less confusing and makes a daemon restart faster (we no longer wait for the stopped daemon to time out) but existing multi-mds configurations may need to be adjusted accordingly to give daemons unique names.
- The 'ceph osd pool delete <poolname>' and 'rados rmpool <poolname>' now have safety interlocks with loud warnings that make you confirm pool removal. Any scripts curenty rely on these functions zapping data without confirmation need to be adjusted accordingly.

NOTABLE CHANGES FROM V0.60

- rbd: incremental backups
- rbd: only set STRIPINGV2 feature if striping parameters are incompatible with old versions
- rbd: require –allow-shrink for resizing images down
- librbd: many bug fixes
- rgw: management REST API
- rgw: fix object corruption on COPY to self
- rgw: new sysvinit script for rpm-based systems
- rgw: allow buckets with '_'
- rgw: CORS support
- mon: many fixes
- mon: improved trimming behavior
- mon: fix data conversion/upgrade problem (from bobtail)
- mon: ability to tune leveldb
- mon: config-keys service to store arbitrary data on monitor
- mon: 'osd crush add|link|unlink|add-bucket ...' commands
- mon: trigger leveldb compaction on trim
- osd: per-rados pool quotas (objects, bytes)
- osd: tool to export, import, and delete PGs from an individual OSD data store
- osd: notify mon on clean shutdown to avoid IO stall
- osd: improved detection of corrupted journals
- osd: ability to tune leveldb
- osd: improve client request throttling
- osd, librados: fixes to the LIST_SNAPS operation
- osd: improvements to scrub error repair
- osd: better prevention of wedging OSDs with ENOSPC
- osd: many small fixes
- mds: fix xattr handling on root inode
- mds: fixed bugs in journal replay
- mds: many fixes
- librados: clean up snapshot constant definitions
- libcephfs: calls to query CRUSH topology (used by Hadoop)
- ceph-fuse, libcephfs: misc fixes to mds session management
- ceph-fuse: disabled cache invalidation (again) due to potential deadlock with kernel
- sysvinit: try to start all daemons despite early failures
- ceph-disk: new 'list' command
- ceph-disk: hotplug fixes for RHEL/CentOS
- ceph-disk: fix creation of OSD data partitions on >2TB disks
- osd: fix udev rules for RHEL/CentOS systems

- fix daemon logging during initial startup

NOTABLE CHANGES FROM V0.56 “BOBTAIL”

- always use installed system leveldb (Gary Lowell)
- auth: ability to require new cephx signatures on messages (still off by default)
- buffer unit testing (Loic Dachary)
- ceph tool: some CLI interface cleanups
- ceph-disk: improve multicluster support, error handling (Sage Weil)
- ceph-disk: support for dm-crypt (Alexandre Marangone)
- ceph-disk: support for sysvinit, directories or partitions (not full disks)
- ceph-disk: fix mkfs args on old distros (Alexandre Marangone)
- ceph-disk: fix creation of OSD data partitions on >2TB disks
- ceph-disk: hotplug fixes for RHEL/CentOS
- ceph-disk: new ‘list’ command
- ceph-fuse, libcephfs: misc fixes to mds session management
- ceph-fuse: disabled cache invalidation (again) due to potential deadlock with kernel
- ceph-fuse: enable kernel cache invalidation (Sam Lang)
- ceph-fuse: fix statfs(2) reporting
- ceph-fuse: session handling cleanup, bug fixes (Sage Weil)
- crush: ability to create, remove rules via CLI
- crush: update weights for all instances of an item, not just the first (Sage Weil)
- fix daemon logging during initial startup
- fixed log rotation (Gary Lowell)
- init-ceph, mkcephfs: close a few security holes with -a (Sage Weil)
- libcephfs: calls to query CRUSH topology (used by Hadoop)
- libcephfs: many fixes, cleanups with the Java bindings
- libcephfs: new topo API requests for Hadoop (Noah Watkins)
- librados: clean up snapshot constant definitions
- librados: fix linger bugs (Josh Durgin)
- librbd: fixed flatten deadlock (Josh Durgin)
- librbd: fixed some locking issues with flatten (Josh Durgin)
- librbd: many bug fixes
- librbd: optionally wait for flush before enabling writeback (Josh Durgin)
- many many cleanups (Danny Al-Gaaf)
- mds, ceph-fuse: fix bugs with replayed requests after MDS restart (Sage Weil)
- mds, ceph-fuse: manage layouts via xattrs
- mds: allow xattrs on root
- mds: fast failover between MDSs (enforce unique mds names)
- mds: fix xattr handling on root inode
- mds: fixed bugs in journal replay
- mds: improve session cleanup (Sage Weil)
- mds: many fixes (Yan Zheng)
- mds: misc bug fixes with clustered MDSs and failure recovery
- mds: misc bug fixes with readdir
- mds: new encoding for all data types (to allow forward/backward compatbility) (Greg Farnum)
- mds: store and update backpointers/traces on directory, file objects (Sam Lang)
- mon: ‘osd crush add|link|unlink|add-bucket ...’ commands
- mon: ability to tune leveldb
- mon: approximate recovery, IO workload stats
- mon: avoid marking entire CRUSH subtrees out (e.g., if an entire rack goes offline)
- mon: config-keys service to store arbitrary data on monitor
- mon: easy adjustment of crush tunables via ‘ceph osd crush tunables ...’
- mon: easy creation of crush rules vai ‘ceph osd rule ...’
- mon: fix data conversion/upgrade problem (from bobtail)
- mon: improved trimming behavior
- mon: many fixes
- mon: new ‘ceph df [detail]’ command
- mon: new checks for identifying and reporting clock drift
- mon: rearchitected to utilize single instance of paxos and a key/value store (Joao Luis)
- mon: safety check for pool deletion
- mon: shut down safely if disk approaches full (Joao Luis)
- mon: trigger leveldb compaction on trim
- msgr: fix comparison of IPv6 addresses (fixes monitor bringup via ceph-deploy, chef)

- msgr: fixed race in connection reset
- msgr: optionally tune TCP buffer size to avoid throughput collapse (Jim Schutt)
- much code cleanup and optimization (Danny Al-Gaaf)
- osd, librados: ability to list watchers (David Zafman)
- osd, librados: fixes to the LIST_SNAPS operation
- osd, librados: new listsnaps command (David Zafman)
- osd: a few journaling bug fixes
- osd: ability to tune leveldb
- osd: add ‘noscrub’, ‘nodeepscrub’ osdmap flags (David Zafman)
- osd: better prevention of wedging OSDs with ENOSPC
- osd: ceph-filestore-dump tool for debugging
- osd: connection handling bug fixes
- osd: deep-scrub omap keys/values
- osd: default to libaio for the journal (some performance boost)
- osd: fix hang in ‘journal aio = true’ mode (Sage Weil)
- osd: fix pg log trimming (avoids memory bloat on degraded clusters)
- osd: fix udev rules for RHEL/CentOS systems
- osd: fixed bug in journal checksums (Sam Just)
- osd: improved client request throttling
- osd: improved handling when disk fills up (David Zafman)
- osd: improved journal corruption detection (Sam Just)
- osd: improved detection of corrupted journals
- osd: improvements to scrub error repair
- osd: make tracking of object snapshot metadata more efficient (Sam Just)
- osd: many small fixes
- osd: misc fixes to PG split (Sam Just)
- osd: move pg info, log into leveldb (== better performance) (David Zafman)
- osd: notify mon on clean shutdown to avoid IO stall
- osd: per-rados pool quotas (objects, bytes)
- osd: refactored watch/notify infrastructure (fixes protocol, removes many bugs) (Sam Just)
- osd: support for improved hashing of PGs across OSDs via HASHPSPOOL pool flag and feature
- osd: tool to export, import, and delete PGs from an individual OSD data store
- osd: trim log more aggressively, avoid appearance of leak memory
- osd: validate snap collections on startup
- osd: verify snap collections on startup (Sam Just)
- radosgw: ACL grants in headers (Caleb Miles)
- radosgw: ability to listen to fastcgi via a port (Guilhem Lettron)
- radosgw: fix object copy onto self (Yehuda Sadeh)
- radosgw: misc fixes
- rbd-fuse: new tool, package
- rbd: avoid FIEMAP when importing from file (it can be buggy)
- rbd: incremental backups
- rbd: only set STRIPINGV2 feature if striping parameters are incompatible with old versions
- rbd: require –allow-shrink for resizing images down
- rbd: udevadm settle on map/unmap to avoid various races (Dan Mick)
- rbd: wait for udev to settle in strategic places (avoid spurious errors, failures)
- rgw: CORS support
- rgw: allow buckets with ‘_’
- rgw: fix Content-Length on 32-bit machines (Jan Harkes)
- rgw: fix log rotation
- rgw: fix object corruption on COPY to self
- rgw: fixed >4MB range requests (Jan Harkes)
- rgw: new sysvinit script for rpm-based systems
- rpm/deb: do not remove /var/lib/ceph on purge (v0.59 was the only release to do so)
- sysvinit: try to start all daemons despite early failures
- upstart: automatically set osd weight based on df (Guilhem Lettron)
- use less memory for logging by default

V0.60

UPGRADING

- Please note that the recently added librados ‘list_snaps’ function call is in a state of flux and is changing slightly in v0.61.

You are advised not to make use of it in v0.59 or v0.60.

NOTABLE CHANGES

- osd: make tracking of object snapshot metadata more efficient (Sam Just)
- osd: misc fixes to PG split (Sam Just)
- osd: improve journal corruption detection (Sam Just)
- osd: improve handling when disk fills up (David Zafman)
- osd: add ‘noscrub’, ‘nodeepscrub’ osdmap flags (David Zafman)
- osd: fix hang in ‘journal aio = true’ mode (Sage Weil)
- ceph-disk-prepare: fix mkfs args on old distros (Alexandre Marangone)
- ceph-disk-activate: improve multicluster support, error handling (Sage Weil)
- librbd: optionally wait for flush before enabling writeback (Josh Durgin)
- crush: update weights for all instances of an item, not just the first (Sage Weil)
- mon: shut down safely if disk approaches full (Joao Luis)
- rgw: fix Content-Length on 32-bit machines (Jan Harkes)
- mds: store and update backpointers/traces on directory, file objects (Sam Lang)
- mds: improve session cleanup (Sage Weil)
- mds, ceph-fuse: fix bugs with replayed requests after MDS restart (Sage Weil)
- ceph-fuse: enable kernel cache invalidation (Sam Lang)
- libcephfs: new topo API requests for Hadoop (Noah Watkins)
- ceph-fuse: session handling cleanup, bug fixes (Sage Weil)
- much code cleanup and optimization (Danny Al-Gaaf)
- use less memory for logging by default
- upstart: automatically set osd weight based on df (Guilhem Lettron)
- init-ceph, mkcephfs: close a few security holes with -a (Sage Weil)
- rpm/deb: do not remove /var/lib/ceph on purge (v0.59 was the only release to do so)

V0.59

UPGRADING

- The monitor is using a completely new storage strategy and intra-cluster protocol. This means that v0.59 and pre-v0.59 monitors do not talk to each other. When you upgrade each one, it will convert its local data store to the new format. Once you upgrade a majority, the quorum will be formed using the new protocol and the old monitors will be blocked out until they too get upgraded. For this reason, we recommend not running a mixed-version cluster for very long.
- ceph-mon now requires the creation of its data directory prior to –mkfs, similarly to what happens on ceph-osd. This directory is no longer automatically created, and custom scripts should be adjusted to reflect just that.

NOTABLE CHANGES

- mon: rearchitected to utilize single instance of paxos and a key/value store (Joao Luis)
- mon: new ‘ceph df [detail]’ command
- osd: support for improved hashing of PGs across OSDs via HASHPSPOOL pool flag and feature
- osd: refactored watch/notify infrastructure (fixes protocol, removes many bugs) (Sam Just)
- osd, librados: ability to list watchers (David Zafman)
- osd, librados: new listsnaps command (David Zafman)
- osd: trim log more aggressively, avoid appearance of leak memory
- osd: misc split fixes
- osd: a few journaling bug fixes
- osd: connection handling bug fixes
- rbd: avoid FIEMAP when importing from file (it can be buggy)
- librados: fix linger bugs (Josh Durgin)
- librbd: fixed flatten deadlock (Josh Durgin)
- rgw: fixed >4MB range requests (Jan Harkes)
- rgw: fix log rotation
- mds: allow xattrs on root
- ceph-fuse: fix statfs(2) reporting
- msgr: optionally tune TCP buffer size to avoid throughput collapse (Jim Schutt)
- consume less memory for logging by default
- always use system leveldb (Gary Lowell)

V0.58

UPGRADING

- The monitor now enforces that MDS names be unique. If you have multiple daemons start with with the same id (e.g., `mds.a`) the second one will implicitly mark the first as failed. This makes things less confusing and makes a daemon restart faster (we no longer wait for the stopped daemon to time out) but existing multi-mds configurations may need to be adjusted accordingly to give daemons unique names.

NOTABLE CHANGES

- librbd: fixed some locking issues with flatten (Josh Durgin)
- rbd: udevadm settle on map/unmap to avoid various races (Dan Mick)
- osd: move pg info, log into leveldb (== better performance) (David Zafman)
- osd: fix pg log trimming (avoids memory bloat on degraded clusters)
- osd: fixed bug in journal checksums (Sam Just)
- osd: verify snap collections on startup (Sam Just)
- ceph-disk-prepare/activate: support for dm-crypt (Alexandre Marangone)
- ceph-disk-prepare/activate: support for sysvinit, directories or partitions (not full disks)
- msgr: fixed race in connection reset
- msgr: fix comparison of IPv6 addresses (fixes monitor bringup via ceph-deploy, chef)
- radosgw: fix object copy onto self (Yehuda Sadeh)
- radosgw: ACL grants in headers (Caleb Miles)
- radosgw: ability to listen to fastcgi via a port (Guilhem Lettron)
- mds: new encoding for all data types (to allow forward/backward compatbility) (Greg Farnum)
- mds: fast failover between MDSs (enforce unique mds names)
- crush: ability to create, remove rules via CLI
- many many cleanups (Danny Al-Gaaf)
- buffer unit testing (Loic Dachary)
- fixed log rotation (Gary Lowell)

V0.57

This development release has a lot of additional functionality accumulated over the last couple months. Most of the bug fixes (with the notable exception of the MDS related work) has already been backported to v0.56.x, and is not mentioned here.

UPGRADING

- The ‘ceph osd pool delete <poolname>’ and ‘rados rmpool <poolname>’ now have safety interlocks with loud warnings that make you confirm pool removal. Any scripts curenty rely on these functions zapping data without confirmation need to be adjusted accordingly.

NOTABLE CHANGES

- osd: default to libaio for the journal (some performance boost)
- osd: validate snap collections on startup
- osd: ceph-filestore-dump tool for debugging
- osd: deep-scrub omap keys/values
- ceph tool: some CLI interface cleanups
- mon: easy adjustment of crush tunables via ‘ceph osd crush tunables ...’
- mon: easy creation of crush rules vai ‘ceph osd rule ...’
- mon: approximate recovery, IO workload stats
- mon: avoid marking entire CRUSH subtrees out (e.g., if an entire rack goes offline)
- mon: safety check for pool deletion
- mon: new checks for identifying and reporting clock drift
- radosgw: misc fixes
- rbd: wait for udev to settle in strategic places (avoid spurious errors, failures)
- rbd-fuse: new tool, package
- mds, ceph-fuse: manage layouts via xattrs
- mds: misc bug fixes with clustered MDSs and failure recovery

- mds: misc bug fixes with readdir
- libcephfs: many fixes, cleanups with the Java bindings
- auth: ability to require new cephx signatures on messages (still off by default)

V0.56.5 “BOBTAIL”

UPGRADING

- ceph-disk[-prepare,-activate] behavior has changed in various ways. There should not be any compatibility issues, but chef users should be aware.

NOTABLE CHANGES

- mon: fix recording of quorum feature set (important for argonaut -> bobtail -> cuttlefish mon upgrades)
- osd: minor peering bug fixes
- osd: fix a few bugs when pools are renamed
- osd: fix occasionally corrupted pg stats
- osd: fix behavior when broken v0.56[.0] clients connect
- rbd: avoid FIEMAP ioctl on import (it is broken on some kernels)
- librbd: fixes for several request/reply ordering bugs
- librbd: only set STRIPINGV2 feature on new images when needed
- librbd: new async flush method to resolve qemu hangs (requires Qemu update as well)
- librbd: a few fixes to flatten
- ceph-disk: support for dm-crypt
- ceph-disk: many backports to allow bobtail deployments with ceph-deploy, chef
- sysvinit: do not stop starting daemons on first failure
- udev: fixed rules for redhat-based distros
- build fixes for raring

For more detailed information, see **the complete changelog**.

V0.56.4 “BOBTAIL”

UPGRADING

- There is a fix in the syntax for the output of ‘ceph osd tree –format=json’.
- The MDS disk format has changed from prior releases *and* from v0.57. In particular, upgrades to v0.56.4 are safe, but you cannot move from v0.56.4 to v0.57 if you are using the MDS for CephFS; you must upgrade directly to v0.58 (or later) instead.

NOTABLE CHANGES

- mon: fix bug in bringup with IPv6
- reduce default memory utilization by internal logging (all daemons)
- rgw: fix for bucket removal
- rgw: reopen logs after log rotation
- rgw: fix multipat upload listing
- rgw: don’t copy object when copied onto self
- osd: fix caps parsing for pools with - or _
- osd: allow pg log trimming when degraded, scrubbing, recoverying (reducing memory consumption)
- osd: fix potential deadlock when ‘journal aio = true’
- osd: various fixes for collection creation/removal, rename, temp collections
- osd: various fixes for PG split
- osd: deep-scrub omap key/value data
- osd: fix rare bug in journal replay
- osd: misc fixes for snapshot tracking
- osd: fix leak in recovery reservations on pool deletion
- osd: fix bug in connection management
- osd: fix for op ordering when rebalancing

- ceph-fuse: report file system size with correct units
- mds: get and set directory layout policies via virtual xattrs
- mds: on-disk format revision (see upgrading note above)
- mkcephfs, init-ceph: close potential security issues with predictable filenames

For more detailed information, see **the complete changelog**.

V0.56.3 "BOBTAIL"

This release has several bug fixes surrounding OSD stability. Most significantly, an issue with OSDs being unresponsive shortly after startup (and occasionally crashing due to an internal heartbeat check) is resolved. Please upgrade.

UPGRADING

- A bug was fixed in which the OSDMap epoch for PGs without any IO requests was not recorded. If there are pools in the cluster that are completely idle (for example, the `data` and `metadata` pools normally used by CephFS), and a large number of OSDMap epochs have elapsed since the `ceph-osd` daemon was last restarted, those maps will get reprocessed when the daemon restarts. This process can take a while if there are a lot of maps. A workaround is to 'touch' any idle pools with IO prior to restarting the daemons after packages are upgraded:

  ```
  rados bench 10 write -t 1 -b 4096 -p {POOLNAME}
  ```

  This will typically generate enough IO to touch every PG in the pool without generating significant cluster load, and also cleans up any temporary objects it creates.

NOTABLE CHANGES

- osd: flush peering work queue prior to start
- osd: persist osdmap epoch for idle PGs
- osd: fix and simplify connection handling for heartbeats
- osd: avoid crash on invalid admin command
- mon: fix rare races with monitor elections and commands
- mon: enforce that OSD reweights be between 0 and 1 (NOTE: not CRUSH weights)
- mon: approximate client, recovery bandwidth logging
- radosgw: fixed some XML formatting to conform to Swift API inconsistency
- radosgw: fix usage accounting bug; add repair tool
- radosgw: make fallback URI configurable (necessary on some web servers)
- librbd: fix handling for interrupted 'unprotect' operations
- mds, ceph-fuse: allow file and directory layouts to be modified via virtual xattrs

For more detailed information, see **the complete changelog**.

V0.56.2 "BOBTAIL"

This release has a wide range of bug fixes, stability improvements, and some performance improvements. Please upgrade.

UPGRADING

- The meaning of the 'osd scrub min interval' and 'osd scrub max interval' has changed slightly. The min interval used to be meaningless, while the max interval would only trigger a scrub if the load was sufficiently low. Now, the min interval option works the way the old max interval did (it will trigger a scrub after this amount of time if the load is low), while the max interval will force a scrub regardless of load. The default options have been adjusted accordingly. If you have customized these in ceph.conf, please review their values when upgrading.
- CRUSH maps that are generated by default when calling `ceph-mon --mkfs` directly now distribute replicas across hosts instead of across OSDs. Any provisioning tools that are being used by Ceph may be affected, although probably for the better, as distributing across hosts is a much more commonly sought behavior. If you use `mkcephfs` to create the cluster, the default CRUSH rule is still inferred by the number of hosts and/or racks in the initial ceph.conf.

NOTABLE CHANGES

- osd: snapshot trimming fixes
- osd: scrub snapshot metadata
- osd: fix osdmap trimming
- osd: misc peering fixes
- osd: stop heartbeating with peers if internal threads are stuck/hung
- osd: PG removal is friendlier to other workloads
- osd: fix recovery start delay (was causing very slow recovery)
- osd: fix scheduling of explicitly requested scrubs
- osd: fix scrub interval config options
- osd: improve recovery vs client io tuning
- osd: improve 'slow request' warning detail for better diagnosis
- osd: default CRUSH map now distributes across hosts, not OSDs
- osd: fix crash on 32-bit hosts triggered by librbd clients
- librbd: fix error handling when talking to older OSDs
- mon: fix a few rare crashes
- ceph command: ability to easily adjust CRUSH tunables
- radosgw: object copy does not copy source ACLs
- rados command: fix omap command usage
- sysvinit script: set ulimit -n properly on remote hosts
- msgr: fix narrow race with message queuing
- fixed compilation on some old distros (e.g., RHEL 5.x)

For more detailed information, see **the complete changelog**.

V0.56.1 "BOBTAIL"

This release has two critical fixes. Please upgrade.

UPGRADING

- There is a protocol compatibility problem between v0.56 and any other version that is now fixed. If your radosgw or RBD clients are running v0.56, they will need to be upgraded too. If they are running a version prior to v0.56, they can be left as is.

NOTABLE CHANGES

- osd: fix commit sequence for XFS, ext4 (or any other non-btrfs) to prevent data loss on power cycle or kernel panic
- osd: fix compatibility for CALL operation
- osd: process old osdmaps prior to joining cluster (fixes slow startup)
- osd: fix a couple of recovery-related crashes
- osd: fix large io requests when journal is in (non-default) aio mode
- log: fix possible deadlock in logging code

For more detailed information, see **the complete changelog**.

V0.56 "BOBTAIL"

Bobtail is the second stable release of Ceph, named in honor of the *Bobtail Squid*: http://en.wikipedia.org/wiki/Bobtail_squid.

KEY FEATURES SINCE V0.48 "ARGONAUT"

- Object Storage Daemon (OSD): improved threading, small-io performance, and performance during recovery
- Object Storage Daemon (OSD): regular "deep" scrubbing of all stored data to detect latent disk errors
- RADOS Block Device (RBD): support for copy-on-write clones of images.
- RADOS Block Device (RBD): better client-side caching.
- RADOS Block Device (RBD): advisory image locking
- Rados Gateway (RGW): support for efficient usage logging/scraping (for billing purposes)
- Rados Gateway (RGW): expanded S3 and Swift API coverage (e.g., POST, multi-object delete)
- Rados Gateway (RGW): improved striping for large objects
- Rados Gateway (RGW): OpenStack Keystone integration
- RPM packages for Fedora, RHEL/CentOS, OpenSUSE, and SLES

- mkcephfs: support for automatically formatting and mounting XFS and ext4 (in addition to btrfs)

UPGRADING

Please refer to the document Upgrading from Argonaut to Bobtail for details.

- Cephx authentication is now enabled by default (since v0.55). Upgrading a cluster without adjusting the Ceph configuration will likely prevent the system from starting up on its own. We recommend first modifying the configuration to indicate that authentication is disabled, and only then upgrading to the latest version.:

```
auth client required = none
auth service required = none
auth cluster required = none
```

- Ceph daemons can be upgraded one-by-one while the cluster is online and in service.
- The `ceph-osd` daemons must be upgraded and restarted *before* any `radosgw` daemons are restarted, as they depend on some new ceph-osd functionality. (The `ceph-mon`, `ceph-osd`, and `ceph-mds` daemons can be upgraded and restarted in any order.)
- Once each individual daemon has been upgraded and restarted, it cannot be downgraded.
- The cluster of `ceph-mon` daemons will migrate to a new internal on-wire protocol once all daemons in the quorum have been upgraded. Upgrading only a majority of the nodes (e.g., two out of three) may expose the cluster to a situation where a single additional failure may compromise availability (because the non-upgraded daemon cannot participate in the new protocol). We recommend not waiting for an extended period of time between `ceph-mon` upgrades.
- The ops log and usage log for radosgw are now off by default. If you need these logs (e.g., for billing purposes), you must enable them explicitly. For logging of all operations to objects in the `.log` pool (see `radosgw-admin log ...`):

```
rgw enable ops log = true
```

  For usage logging of aggregated bandwidth usage (see `radosgw-admin usage ...`):

```
rgw enable usage log = true
```

- You should not create or use “format 2” RBD images until after all `ceph-osd` daemons have been upgraded. Note that “format 1” is still the default. You can use the new `ceph osd ls` and `ceph tell osd.N version` commands to doublecheck your cluster. `ceph osd ls` will give a list of all OSD IDs that are part of the cluster, and you can use that to write a simple shell loop to display all the OSD version strings:

```
for i in $(ceph osd ls); do
    ceph tell osd.${i} version
done
```

COMPATIBILITY CHANGES

- The ‘ceph osd create [<uuid>]’ command now rejects an argument that is not a UUID. (Previously it would take take an optional integer OSD id.) This correct syntax has been ‘ceph osd create [<uuid>]’ since v0.47, but the older calling convention was being silently ignored.
- The CRUSH map root nodes now have type `root` instead of type `pool`. This avoids confusion with RADOS pools, which are not directly related. Any scripts or tools that use the `ceph osd crush ...` commands may need to be adjusted accordingly.
- The `ceph osd pool create <poolname> <pgnum>` command now requires the `pgnum` argument. Previously this was optional, and would default to 8, which was almost never a good number.
- Degraded mode (when there fewer than the desired number of replicas) is now more configurable on a per-pool basis, with the min_size parameter. By default, with min_size 0, this allows I/O to objects with N - floor(N/2) replicas, where N is the total number of expected copies. Argonaut behavior was equivalent to having min_size = 1, so I/O would always be possible if any completely up to date copy remained. min_size = 1 could result in lower overall availability in certain

cases, such as flapping network partitions.

- The sysvinit start/stop script now defaults to adjusting the max open files ulimit to 16384. On most systems the default is 1024, so this is an increase and won't break anything. If some system has a higher initial value, however, this change will lower the limit. The value can be adjusted explicitly by adding an entry to the `ceph.conf` file in the appropriate section. For example:

```
[global]
        max open files = 32768
```

- 'rbd lock list' and 'rbd showmapped' no longer use tabs as separators in their output.
- There is configurable limit on the number of PGs when creating a new pool, to prevent a user from accidentally specifying a ridiculous number for pg_num. It can be adjusted via the 'mon max pool pg num' option on the monitor, and defaults to 65536 (the current max supported by the Linux kernel client).
- The osd capabilities associated with a rados user have changed syntax since 0.48 argonaut. The new format is mostly backwards compatible, but there are two backwards-incompatible changes:
  - specifying a list of pools in one grant, i.e. 'allow r pool=foo,bar' is now done in separate grants, i.e. 'allow r pool=foo, allow r pool=bar'.
  - restricting pool access by pool owner ('allow r uid=foo') is removed. This feature was not very useful and unused in practice.

  The new format is documented in the ceph-authtool man page.
- 'rbd cp' and 'rbd rename' use rbd as the default destination pool, regardless of what pool the source image is in. Previously they would default to the same pool as the source image.
- 'rbd export' no longer prints a message for each object written. It just reports percent complete like other long-lasting operations.
- 'ceph osd tree' now uses 4 decimal places for weight so output is nicer for humans
- Several monitor operations are now idempotent:
  - ceph osd pool create
  - ceph osd pool delete
  - ceph osd pool mksnap
  - ceph osd rm
  - ceph pg <pgid> revert

NOTABLE CHANGES

- auth: enable cephx by default
- auth: expanded authentication settings for greater flexibility
- auth: sign messages when using cephx
- build fixes for Fedora 18, CentOS/RHEL 6
- ceph: new 'osd ls' and 'osd tell <osd.N> version' commands
- ceph-debugpack: misc improvements
- ceph-disk-prepare: creates and labels GPT partitions
- ceph-disk-prepare: support for external journals, default mount/mkfs options, etc.
- ceph-fuse/libcephfs: many misc fixes, admin socket debugging
- ceph-fuse: fix handling for .. in root directory
- ceph-fuse: many fixes (including memory leaks, hangs)
- ceph-fuse: mount helper (mount.fuse.ceph) for use with /etc/fstab
- ceph.spec: misc packaging fixes
- common: thread pool sizes can now be adjusted at runtime
- config: $pid is now available as a metavariable
- crush: default root of tree type is now 'root' instead of 'pool' (to avoid confusiong wrt rados pools)
- crush: fixed retry behavior with chooseleaf via tunable
- crush: tunables documented; feature bit now present and enforced
- libcephfs: java wrapper
- librados: several bug fixes (rare races, locking errors)
- librados: some locking fixes
- librados: watch/notify fixes, misc memory leaks

- librbd: a few fixes to ‘discard’ support
- librbd: fine-grained striping feature
- librbd: fixed memory leaks
- librbd: fully functional and documented image cloning
- librbd: image (advisory) locking
- librbd: improved caching (of object non-existence)
- librbd: ‘flatten’ command to sever clone parent relationship
- librbd: ‘protect’/’unprotect’ commands to prevent clone parent from being deleted
- librbd: clip requests past end-of-image.
- librbd: fixes an issue with some windows guests running in qemu (remove floating point usage)
- log: fix in-memory buffering behavior (to only write log messages on crash)
- mds: fix ino release on abort session close, relative getattr path, mds shutdown, other misc items
- mds: misc fixes
- mkcephfs: fix for default keyring, osd data/journal locations
- mkcephfs: support for formatting xfs, ext4 (as well as btrfs)
- init: support for automatically mounting xfs and ext4 osd data directories
- mon, radosgw, ceph-fuse: fixed memory leaks
- mon: improved ENOSPC, fs error checking
- mon: less-destructive ceph-mon –mkfs behavior
- mon: misc fixes
- mon: more informative info about stuck PGs in ‘health detail’
- mon: information about recovery and backfill in ‘pg <pgid> query’
- mon: new ‘osd crush create-or-move ...’ command
- mon: new ‘osd crush move ...’ command lets you rearrange your CRUSH hierarchy
- mon: optionally dump ‘osd tree’ in json
- mon: configurable cap on maximum osd number (mon max osd)
- mon: many bug fixes (various races causing ceph-mon crashes)
- mon: new on-disk metadata to facilitate future mon changes (post-bobtail)
- mon: election bug fixes
- mon: throttle client messages (limit memory consumption)
- mon: throttle osd flapping based on osd history (limits osdmap ´thrashing’ on overloaded or unhappy clusters)
- mon: ‘report’ command for dumping detailed cluster status (e.g., for use when reporting bugs)
- mon: osdmap flags like noup, noin now cause a health warning
- msgr: improved failure handling code
- msgr: many bug fixes
- osd, mon: honor new ‘nobackfill’ and ‘norecover’ osdmap flags
- osd, mon: use feature bits to lock out clients lacking CRUSH tunables when they are in use
- osd: backfill reservation framework (to avoid flooding new osds with backfill data)
- osd: backfill target reservations (improve performance during recovery)
- osd: better tracking of recent slow operations
- osd: capability grammar improvements, bug fixes
- osd: client vs recovery io prioritization
- osd: crush performance improvements
- osd: default journal size to 5 GB
- osd: experimental support for PG “splitting” (pg_num adjustment for existing pools)
- osd: fix memory leak on certain error paths
- osd: fixed detection of EIO errors from fs on read
- osd: major refactor of PG peering and threading
- osd: many bug fixes
- osd: more/better dump info about in-progress operations
- osd: new caps structure (see compatibility notes)
- osd: new ‘deep scrub’ will compare object content across replicas (once per week by default)
- osd: new ‘lock’ rados class for generic object locking
- osd: optional ‘min’ pg size
- osd: recovery reservations
- osd: scrub efficiency improvement
- osd: several out of order reply bug fixes
- osd: several rare peering cases fixed
- osd: some performance improvements related to request queuing
- osd: use entire device if journal is a block device
- osd: use syncfs(2) when kernel supports it, even if glibc does not
- osd: various fixes for out-of-order op replies
- rados: ability to copy, rename pools
- rados: bench command now cleans up after itself
- rados: ‘cppool’ command to copy rados pools

- rados: ‘rm’ now accepts a list of objects to be removed
- radosgw: POST support
- radosgw: REST API for managing usage stats
- radosgw: fix bug in bucket stat updates
- radosgw: fix copy-object vs attributes
- radosgw: fix range header for large objects, ETag quoting, GMT dates, other compatibility fixes
- radosgw: improved garbage collection framework
- radosgw: many small fixes, cleanups
- radosgw: openstack keystone integration
- radosgw: stripe large (non-multipart) objects
- radosgw: support for multi-object deletes
- radosgw: support for swift manifest objects
- radosgw: vanity bucket dns names
- radosgw: various API compatibility fixes
- rbd: import from stdin, export to stdout
- rbd: new ‘ls -l’ option to view images with metadata
- rbd: use generic id and keyring options for ‘rbd map’
- rbd: don’t issue usage on errors
- udev: fix symlink creation for rbd images containing partitions
- upstart: job files for all daemon types (not enabled by default)
- wireshark: ceph protocol dissector patch updated

V0.54

UPGRADING

- The osd capabilities associated with a rados user have changed syntax since 0.48 argonaut. The new format is mostly backwards compatible, but there are two backwards-incompatible changes:
    - specifying a list of pools in one grant, i.e. ‘allow r pool=foo,bar’ is now done in separate grants, i.e. ‘allow r pool=foo, allow r pool=bar’.
    - restricting pool access by pool owner (‘allow r uid=foo’) is removed. This feature was not very useful and unused in practice.

  The new format is documented in the ceph-authtool man page.
- Bug fixes to the new osd capability format parsing properly validate the allowed operations. If an existing rados user gets permissions errors after upgrading, its capabilities were probably misconfigured. See the ceph-authtool man page for details on osd capabilities.
- ‘rbd lock list’ and ‘rbd showmapped’ no longer use tabs as separators in their output.

V0.48.3 “ARGONAUT”

This release contains a critical fix that can prevent data loss or corruption after a power loss or kernel panic event. Please upgrade immediately.

UPGRADING

- If you are using the undocumented `ceph-disk-prepare` and `ceph-disk-activate` tools, they have several new features and some additional functionality. Please review the changes in behavior carefully before upgrading.
- The .deb packages now require xfsprogs.

NOTABLE CHANGES

- filestore: fix op_seq write order (fixes journal replay after power loss)
- osd: fix occasional indefinitely hung “slow” request
- osd: fix encoding for pool_snap_info_t when talking to pre-v0.48 clients
- osd: fix heartbeat check
- osd: reduce log noise about rbd watch
- log: fixes for deadlocks in the internal logging code

- log: make log buffer size adjustable
- init script: fix for ‘ceph status’ across machines
- radosgw: fix swift error handling
- radosgw: fix swift authentication concurrency bug
- radosgw: don’t cache large objects
- radosgw: fix some memory leaks
- radosgw: fix timezone conversion on read
- radosgw: relax date format restrictions
- radosgw: fix multipart overwrite
- radosgw: stop processing requests on client disconnect
- radosgw: avoid adding port to url that already has a port
- radosgw: fix copy to not override ETAG
- common: make parsing of ip address lists more forgiving
- common: fix admin socket compatibility with old protocol (for collectd plugin)
- mon: drop dup commands on paxos reset
- mds: fix loner selection for multiclient workloads
- mds: fix compat bit checks
- ceph-fuse: fix segfault on startup when keyring is missing
- ceph-authtool: fix usage
- ceph-disk-activate: misc backports
- ceph-disk-prepare: misc backports
- debian: depend on xfsprogs (we use xfs by default)
- rpm: build rpms, some related Makefile changes

For more detailed information, see **the complete changelog**.

V0.48.2 “ARGONAUT”

UPGRADING

- The default search path for keyring files now includes /etc/ceph/ceph.$name.keyring. If such files are present on your cluster, be aware that by default they may now be used.
- There are several changes to the upstart init files. These have not been previously documented or recommended. Any existing users should review the changes before upgrading.
- The ceph-disk-prepare and ceph-disk-active scripts have been updated significantly. These have not been previously documented or recommended. Any existing users should review the changes before upgrading.

NOTABLE CHANGES

- mkcephfs: fix keyring generation for mds, osd when default paths are used
- radosgw: fix bug causing occasional corruption of per-bucket stats
- radosgw: workaround to avoid previously corrupted stats from going negative
- radosgw: fix bug in usage stats reporting on busy buckets
- radosgw: fix Content-Range: header for objects bigger than 2 GB.
- rbd: avoid leaving watch acting when command line tool errors out (avoids 30s delay on subsequent operations)
- rbd: friendlier use of –pool/–image options for import (old calling convention still works)
- librbd: fix rare snapshot creation race (could “lose” a snap when creation is concurrent)
- librbd: fix discard handling when spanning holes
- librbd: fix memory leak on discard when caching is enabled
- objecter: misc fixes for op reordering
- objecter: fix for rare startup-time deadlock waiting for osdmap
- ceph: fix usage
- mon: reduce log noise about “check_sub”
- ceph-disk-activate: misc fixes, improvements
- ceph-disk-prepare: partition and format osd disks automatically
- upstart: start everyone on a reboot
- upstart: always update the osd crush location on start if specified in the config
- config: add /etc/ceph/ceph.$name.keyring to default keyring search path
- ceph.spec: don’t package crush headers

For more detailed information, see **the complete changelog**.

V0.48.1 “ARGONAUT”

UPGRADING

- The radosgw usage trim function was effectively broken in v0.48. Earlier it would remove more usage data than what was requested. This is fixed in v0.48.1, but the fix is incompatible. The v0.48 radosgw-admin tool cannot be used to initiate the trimming; please use the v0.48.1 version.
- v0.48.1 now explicitly indicates support for the CRUSH_TUNABLES feature. No other version of Ceph requires this, yet, but future versions will when the tunables are adjusted from their historical defaults.
- There are no other compatibility changes between v0.48.1 and v0.48.

NOTABLE CHANGES

- mkcephfs: use default ‘keyring’, ‘osd data’, ‘osd journal’ paths when not specified in conf
- msgr: various fixes to socket error handling
- osd: reduce scrub overhead
- osd: misc peering fixes (past_interval sharing, pgs stuck in ‘peering’ states)
- osd: fail on EIO in read path (do not silently ignore read errors from failing disks)
- osd: avoid internal heartbeat errors by breaking some large transactions into pieces
- osd: fix osdmap catch-up during startup (catch up and then add daemon to osdmap)
- osd: fix spurious ‘misdirected op’ messages
- osd: report scrub status via ‘pg ... query’
- rbd: fix race when watch registrations are resent
- rbd: fix rbd image id assignment scheme (new image data objects have slightly different names)
- rbd: fix perf stats for cache hit rate
- rbd tool: fix off-by-one in key name (crash when empty key specified)
- rbd: more robust udev rules
- rados tool: copy object, pool commands
- radosgw: fix in usage stats trimming
- radosgw: misc API compatibility fixes (date strings, ETag quoting, swift headers, etc.)
- ceph-fuse: fix locking in read/write paths
- mon: fix rare race corrupting on-disk data
- config: fix admin socket ‘config set’ command
- log: fix in-memory log event gathering
- debian: remove crush headers, include librados-config
- rpm: add ceph-disk-{activate, prepare}

For more detailed information, see **the complete changelog**.

V0.48 “ARGONAUT”

UPGRADING

- This release includes a disk format upgrade. Each ceph-osd daemon, upon startup, will migrate its locally stored data to the new format. This process can take a while (for large object counts, even hours), especially on non-btrfs file systems.

- To keep the cluster available while the upgrade is in progress, we recommend you upgrade a storage node or rack at a time, and wait for the cluster to recover each time. To prevent the cluster from moving data around in response to the OSD daemons being down for minutes or hours, you may want to:

  ```
  ceph osd set noout
  ```

  This will prevent the cluster from marking down OSDs as “out” and re-replicating the data elsewhere. If you do this, be sure to clear the flag when the upgrade is complete:

  ```
  ceph osd unset noout
  ```

- There is a encoding format change internal to the monitor cluster. The monitor daemons are careful to switch to the new

format only when all members of the quorum support it. However, that means that a partial quorum with new code may move to the new format, and a recovering monitor running old code will be unable to join (it will crash). If this occurs, simply upgrading the remaining monitor will resolve the problem.

- The ceph tool's -s and -w commands from previous versions are incompatible with this version. Upgrade your client tools at the same time you upgrade the monitors if you rely on those commands.
- It is not possible to downgrade from v0.48 to a previous version.

NOTABLE CHANGES

- osd: stability improvements
- osd: capability model simplification
- osd: simpler/safer –mkfs (no longer removes all files; safe to re-run on active osd)
- osd: potentially buggy FIEMAP behavior disabled by default
- rbd: caching improvements
- rbd: improved instrumentation
- rbd: bug fixes
- radosgw: new, scalable usage logging infrastructure
- radosgw: per-user bucket limits
- mon: streamlined process for setting up authentication keys
- mon: stability improvements
- mon: log message throttling
- doc: improved documentation (ceph, rbd, radosgw, chef, etc.)
- config: new default locations for daemon keyrings
- config: arbitrary variable substitutions
- improved 'admin socket' daemon admin interface (ceph –admin-daemon ...)
- chef: support for multiple monitor clusters
- upstart: basic support for monitors, mds, radosgw; osd support still a work in progress.

The new default keyring locations mean that when enabling authentication (`auth supported = cephx`), keyring locations do not need to be specified if the keyring file is located inside the daemon's data directory (`/var/lib/ceph/$type/ceph-$id` by default).

There is also a lot of librbd code in this release that is laying the groundwork for the upcoming layering functionality, but is not actually used. Likewise, the upstart support is still incomplete and not recommended; we will backport that functionality later if it turns out to be non-disruptive.