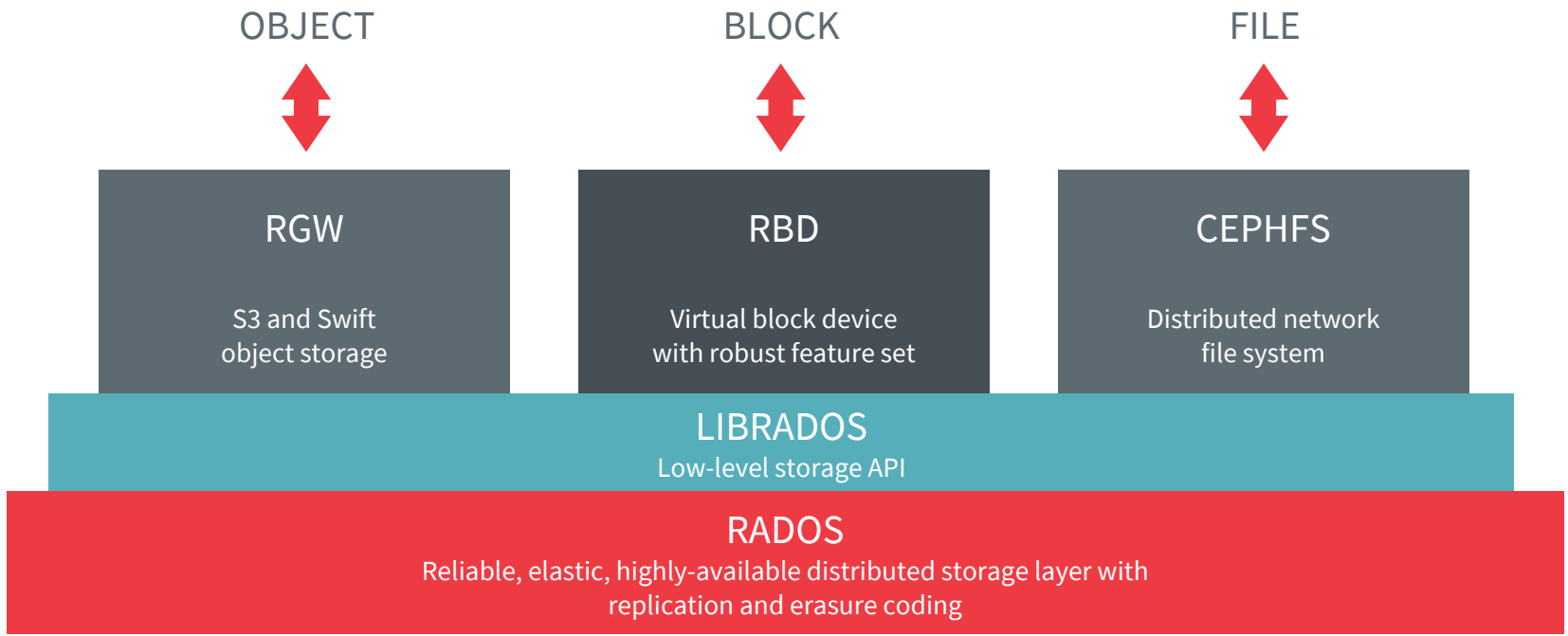
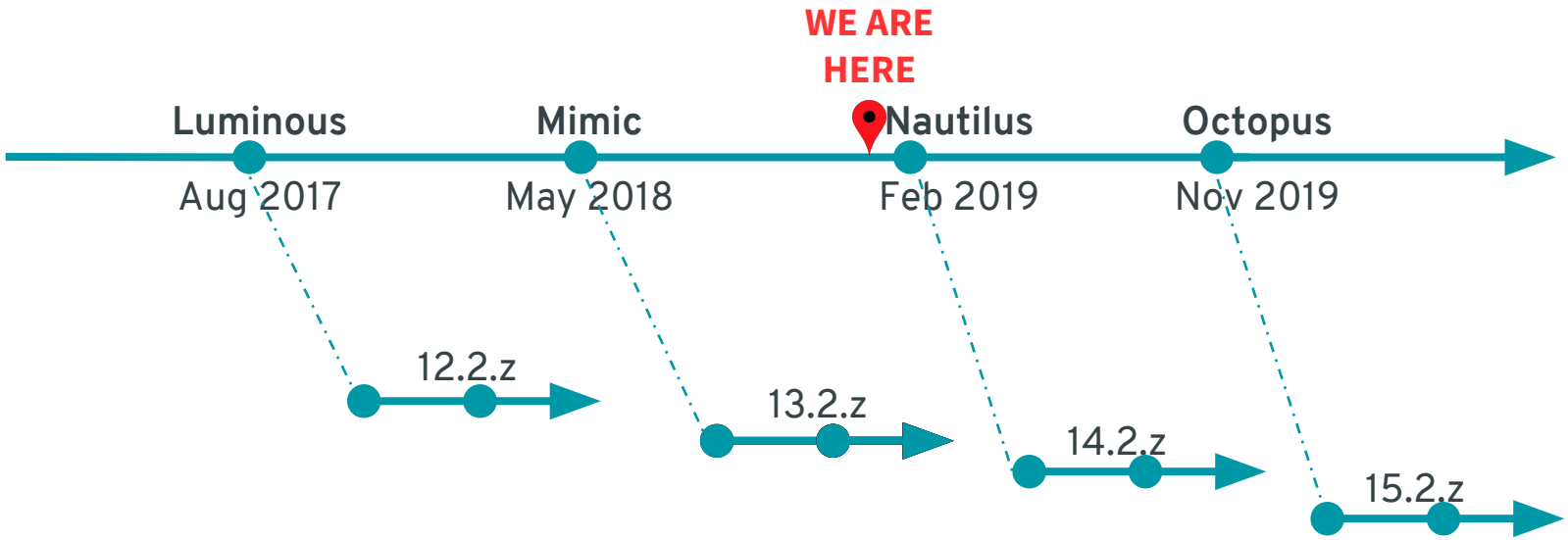
WHAT'S NEW IN CEPH
NAUTILUS

Sage Weil - Red Hat
FOSDEM - 2019.02.03

CEPH UNIFIED STORAGE PLATFORM



RELEASE SCHEDULE



- Stable, named release every 9 months
- Backports for 2 releases
- Upgrade up to 2 releases at a time
  - (e.g., Luminous → Nautilus, Mimic → Octopus)

FOUR CEPH PRIORITIES

Usability and management

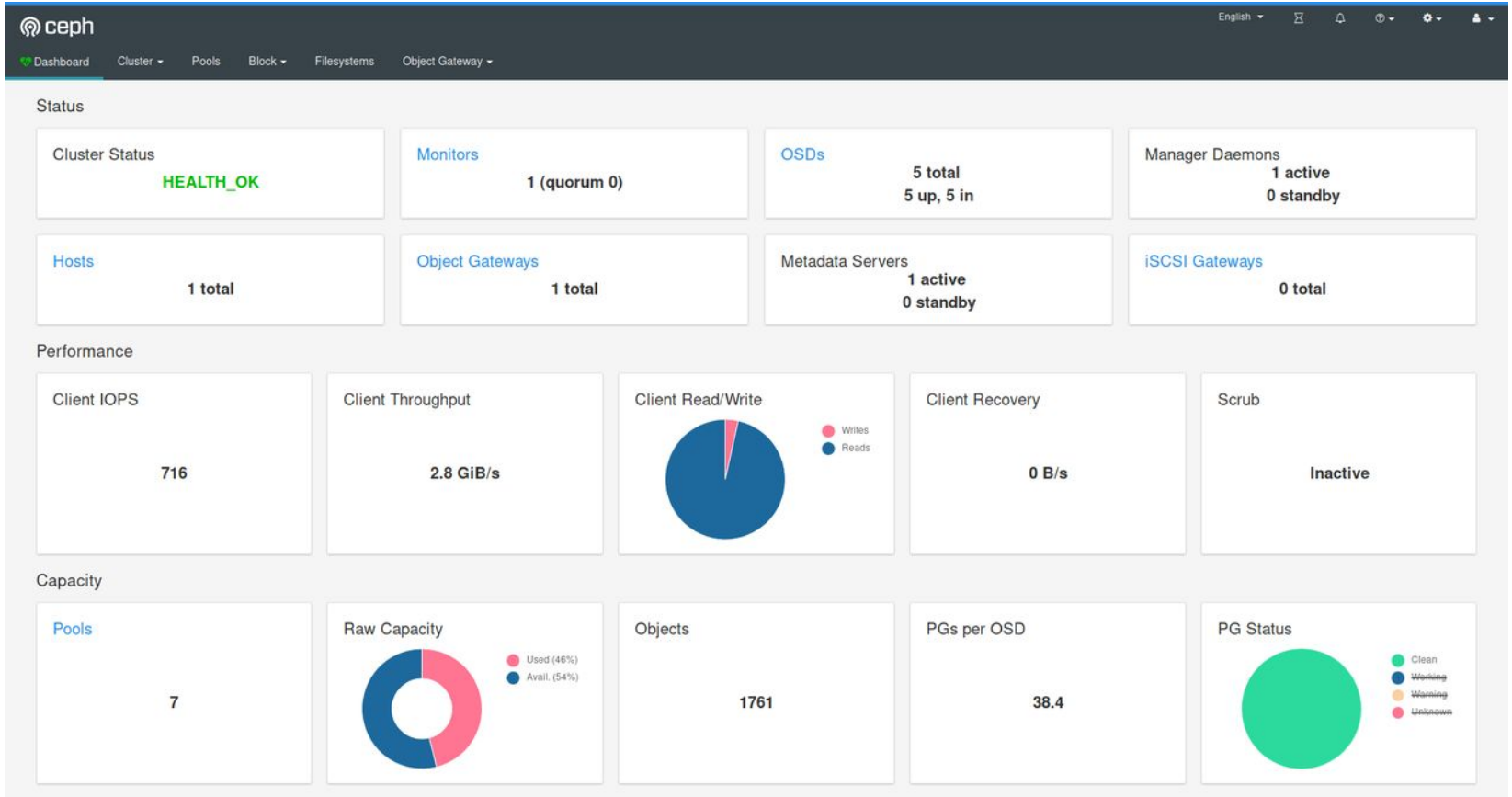
Container ecosystem

Performance

Multi- and hybrid cloud

EASE OF USE AND MANAGEMENT

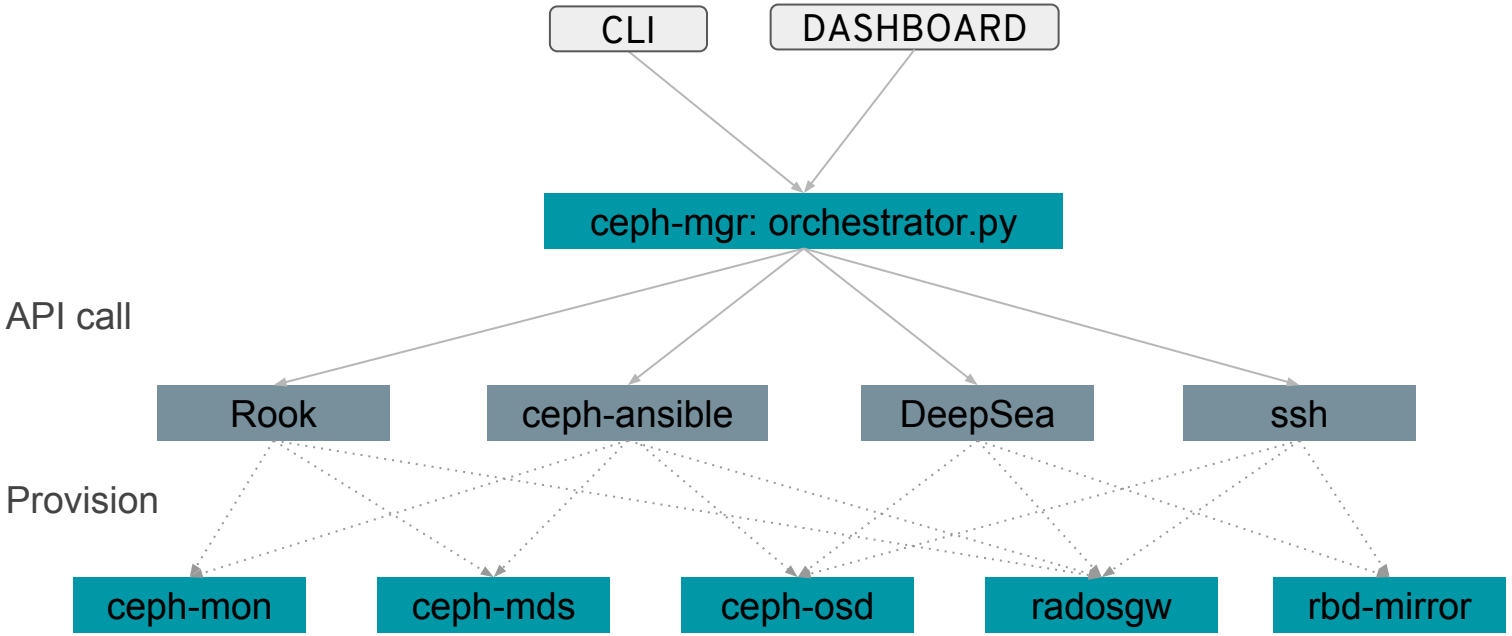
DASHBOARD


ceph
English
Dashboard
Cluster
Pools
Block
Filesystems
Object Gateway
Status
Cluster Status
HEALTH_OK
Monitors
1 (quorum 0)
OSDs
5 total
5 up, 5 in
Manager Daemons
1 active
0 standby
Hosts
1 total
Object Gateways
1 total
Metadata Servers
1 active
0 standby
iSCSI Gateways
0 total
Performance
Client IOPS
716
Client Throughput
2.8 GiB/s
Client Read/Write
Writes
Reads
Client Recovery
0 B/s
Scrub
Inactive
Capacity
Pools
7
Raw Capacity
Used (46%)
Avail. (54%)
Objects
1761
PGs per OSD
38.4
PG Status
Clean
Working
Warning
Unknown

DASHBOARD

- Community convergence in single built-in dashboard
  - Based on SUSE's OpenATTIC and our dashboard prototype
  - SUSE (~10 ppl), Red Hat (~3 ppl), misc community contributors
  - (Finally!)
- Built-in and self-hosted
  - Trivial deployment, tightly integrated with ceph-mgr
  - Easily skinned, localization in progress
- Management functions
  - RADOS, RGW, RBD, CephFS
- Metrics and monitoring
  - Integrates grafana dashboards from ceph-metrics
- Hardware/deployment management in progress...

ORCHESTRATOR SANDWICH



ORCHESTRATOR SANDWICH

- Abstract deployment functions
  - Fetching node inventory
  - Creating or destroying daemon deployments
  - Blinking device LEDs
- **Unified CLI** for managing Ceph daemons
  - `ceph orchestrator device ls [node]`
  - `ceph orchestrator osd create [flags] node device [device]`
  - `ceph orchestrator mon rm [name]`
  - ...
- Enable **dashboard GUI** for deploying and managing daemons
  - Coming post-Nautilus, but some basics are likely to be backported
- Nautilus includes framework and partial implementation

PG AUTOSCALING

- Picking pg_num has historically been “black magic”
  - Limited/confusing guidance on what value(s) to choose
  - pg_num could be increased, but never decreased
- Nautilus: pg_num can be reduced
- Nautilus: pg_num can be automagically tuned in the background
  - Based on usage (how much data in each pool)
  - Administrator can optionally hint about future/expected usage
  - Ceph can either issue health warning or initiate changes itself

```
$ ceph osd pool autoscale-status
POOL      SIZE   TARGET SIZE  RATE  RAW CAPACITY   RATIO  TARGET RATIO  PG_NUM  NEW PG_NUM  AUTOSCALE
   a    12900M                3.0        82431M  0.4695                       8         128  warn
   c         0                3.0        82431M  0.0000        0.2000         1          64  warn
   b         0        953.6M  3.0        82431M  0.0347                       8              warn
```

DEVICE HEALTH METRICS

- OSD and mon report underlying storage devices, scrape SMART metrics

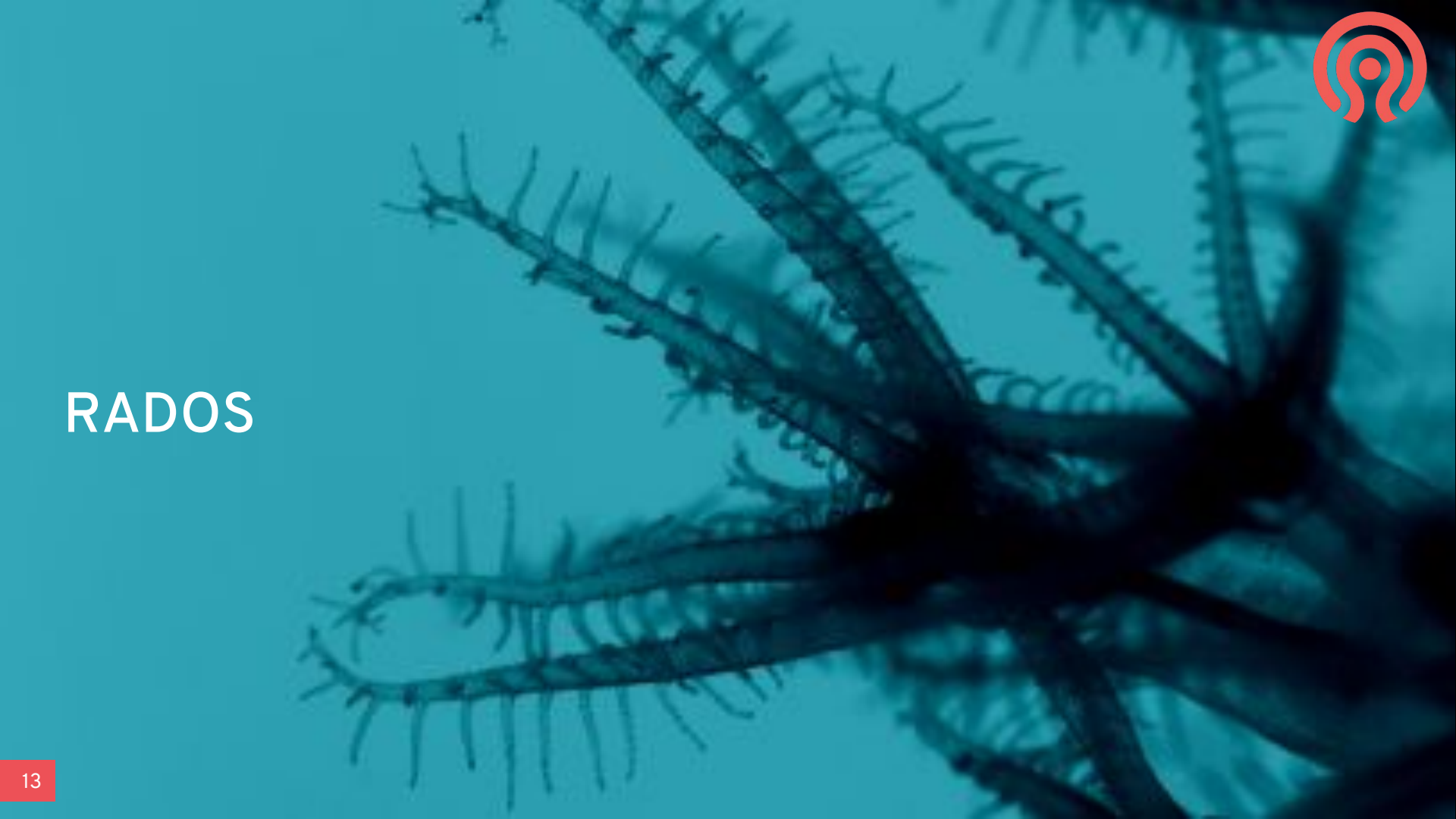
```
# ceph device ls
DEVICE                                     HOST:DEV      DAEMONS    LIFE EXPECTANCY
Crucial_CT1024M550SSD1_14160C164100        stud:sdd      osd.40     >5w
Crucial_CT1024M550SSD1_14210C25EB65        cpach:sde     osd.18     >5w
Crucial_CT1024M550SSD1_14210C25F936        stud:sde      osd.41     >8d
INTEL_SSDPE2ME400G4_CVMD5442003M400FGN     cpach:nvme1n1 osd.10
INTEL_SSDPE2MX012T4_CVPD6185002R1P2QGN     stud:nvme0n1  osd.1
ST2000NX0253_S4608PDF                      cpach:sdo     osd.7
ST2000NX0253_S460971P                      cpach:sdn     osd.8
Samsung_SSD_850_EVO_1TB_S2RENX0J500066T    cpach:sdb     mon.cpach  >5w
```

- Failure prediction
  - Local mode: pretrained model in ceph-mgr predicts remaining life
  - Cloud mode: SaaS based service (free or paid) from ProphetStor
- Optional automatic mitigation
  - Raise health alerts (about specific failing devices, or looming failure storm)
  - Automatically mark soon-to-fail OSDs “out”

CRASH REPORTS

- Previously crashes would manifest as a splat in a daemon log file, usually unnoticed...
- Now concise crash reports logged to /var/lib/ceph/crash/
  - Daemon, timestamp, version
  - Stack trace
- Reports are regularly posted to the mon/mgr
- ‘ceph crash ls’, ‘ceph crash info <id>’, ...
- If user opts in, telemetry module can phone home crashes to Ceph devs

RADOS



MSGR2

- New version of the Ceph on-wire protocol
- Goodness
  - **Encryption on the wire**
  - Improved feature negotiation
  - Improved support for extensible authentication
    - Kerberos is coming soon… hopefully in Octopus!
  - Infrastructure to support dual stack IPv4 and IPv6 (not quite complete)
- Move to IANA-assigned monitor port 3300
- Dual support for v1 and v2 protocols
  - After upgrade, monitor will start listening on 3300, other daemons will starting binding to new v2 ports
  - Kernel support for v2 will come later

RADOS - MISC MANAGEMENT

- osd_target_memory
  - Set target memory usage and OSD caches auto-adjust to fit
- NUMA management, pinning
  - ‘ceph osd numa-status’ to see OSD network and storage NUMA node
  - ‘ceph config set osd.<osd-id> osd_numa_node <num> ; ceph osd down <osd-id>’
- Improvements to centralized config mgmt
  - Especially options from mgr modules
  - Type checking, live changes without restarting ceph-mgr
- Progress bars on recovery, etc.
  - ‘ceph progress’
  - Eventually this will get rolled into ‘ceph -s’...
- ‘Misplaced’ is no longer HEALTH_WARN

BLUESTORE IMPROVEMENTS

- New ‘bitmap’ allocator
  - Faster
  - Predictable and low memory utilization (~10MB RAM per TB SDD, ~3MB RAM per TB HDD)
  - Less fragmentation
- Intelligent cache management
  - Balance memory allocation between RocksDB cache, BlueStore onodes, data
- Per-pool utilization metrics
  - User data, allocated space, compressed size before/after, omap space consumption
  - These bubble up to ‘ceph df’ to monitor e.g., effectiveness of compression
- Misc performance improvements

RADOS MISCELLANY

- CRUSH can convert/reclassify legacy maps
  - Transition from old, hand-crafted maps to new device classes (new in Luminous) no longer shuffles all data
- OSD hard limit on PG log length
  - Avoids corner cases that could cause OSD memory utilization to grow unbounded
- Clay erasure code plugin
  - Better recovery efficiency when <m nodes fail (for a k+m code)

RGW

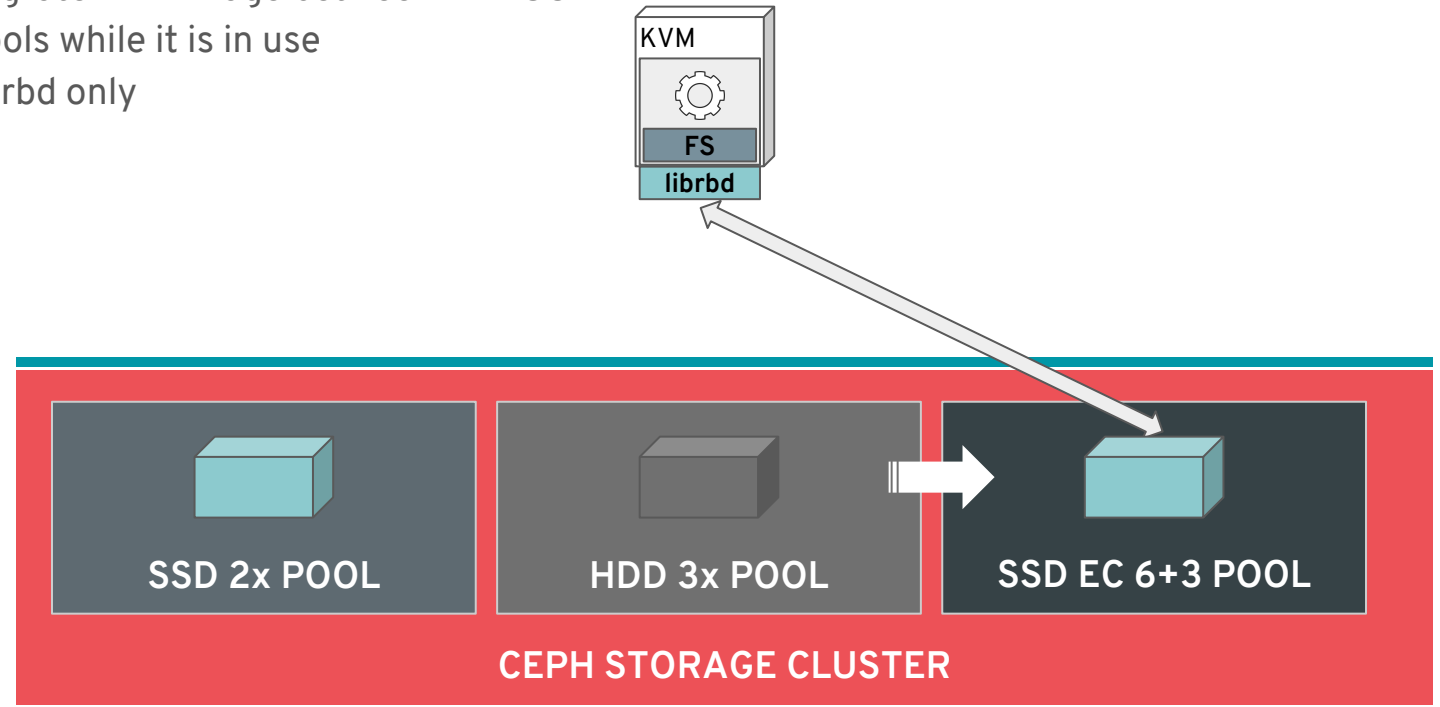
RGW

- pub/sub
  - Subscribe to events like PUT
  - Polling interface, recently demoed with knative at KubeCon Seattle
  - Push interface to AMQ, Kafka coming soon
- Archive zone
  - Enable bucket versioning and retain all copies of all objects
- Tiering policy, lifecycle management
  - Implements S3 API for tiering and expiration
- Beast frontend for RGW
  - Based on boost::asio
  - Better performance and efficiency
- STS

RBD

RBD LIVE IMAGE MIGRATION

- Migrate RBD image between RADOS pools while it is in use
- librbd only



RBD TOP

- RADOS infrastructure
  - ceph-mgr instructs OSDs to sample requests
    - Optionally with some filtering by pool, object name, client, etc.
  - Results aggregated by mgr
- rbd CLI presents this for RBD images specifically

RBD MISC

- rbd-mirror: remote cluster endpoint config stored in cluster
  - Simpler configuration experience!
- Namespace support
  - Lock down tenants to a slice of a pool
  - Private view of images, etc.
- Pool-level config overrides
  - Simpler configuration
- Creation, access, modification timestamps

CEPHFS

CEPHFS VOLUMES AND SUBVOLUMES

- Multi-fs (“volume”) support stable
  - Each CephFS volume has independent set of RADOS pools, MDS cluster
- First-class subvolume concept
  - Sub-directory of a volume with quota, unique cephx user, and restricted to a RADOS namespace
  - Based on ceph_volume_client.py, written for OpenStack Manila driver, now part of ceph-mgr
- ‘ceph fs volume …’, ‘ceph fs subvolume …’

CEPHFS NFS GATEWAYS

- Clustered nfs-ganesha
  - active/active
  - Correct failover semantics (i.e., managed NFS grace period)
  - nfs-ganesha daemons use RADOS for configuration, grace period state
  - (See Jeff Layton's devconf.cz talk recording)
- nfs-ganesha daemons fully managed via new orchestrator interface
  - Fully supported with Rook; others to follow
  - Full support from CLI to Dashboard
- Mapped to new volume/subvolume concept

CEPHFS MISC

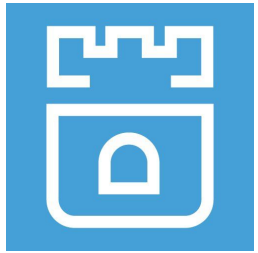
- Cephfs shell
  - CLI tool with shell-like commands (cd, ls, mkdir, rm)
  - Easily scripted
  - Useful for e.g., setting quota attributes on directories without mounting the fs
- Performance, MDS scale(-up) improvements
  - Many fixes for MDSs with large amounts of RAM
  - MDS balancing improvements for multi-MDS clusters

CONTAINER ECOSYSTEM

KUBERNETES

- Expose Ceph storage to Kubernetes
  - Any scale-out infrastructure platform needs scale-out storage
- Run Ceph clusters in Kubernetes
  - Simplify/hide OS dependencies
  - Finer control over upgrades
  - Schedule deployment of Ceph daemons across hardware nodes
- Kubernetes as “distributed OS”

ROOK



- All-in on Rook as a robust operator for Ceph in Kubernetes
  - Extremely easy to get Ceph up and running!
- Intelligent management of Ceph daemons
  - add/remove monitors while maintaining quorum
  - Schedule stateless daemons (rgw, nfs, rbd-mirror) across nodes
- Kubernetes-style provisioning of storage
  - Persistent Volumes (RWO and RWX)
  - Coming: dynamic provisioning of RGW users and buckets
- Enthusiastic user community, CNCF incubation project
- Working hard toward v1.0 release
  - Focus on ability to support in production environments

BAREBONES CONTAINER ORCHESTRATION

- We have: rook, deepsea, ansible, and ssh orchestrator (WIP) implementations
- ssh orch gives mgr a root ssh key to Ceph nodes
  - Moral equivalent/successor of ceph-deploy, but built into the mgr
  - Plan is to eventually combine with a ceph-bootstrap.sh that starts mon+mgr on current host
- ceph-ansible can run daemons in containers
  - Creates a systemd unit file for each daemon that does ‘docker run …’
- Plan to teach ssh orchestrator to do the same
  - Easier install
    - s/fiddling with $random_distro repos/choose container registry and image/
  - Daemons can be upgraded individually, in any order, instead of by host

COMMUNITY

THE CEPH FOUNDATION: WHAT IS IT?

- Organized as a directed fund under the Linux Foundation
  - Members contribute and pool funds
  - Governing board manages expenditures
- Tasked with supporting the Ceph project community
  - **Financial support** for project infrastructure, events, internships, outreach, marketing, and related efforts
  - **Forum for coordinating activities** and investments, providing guidance to technical teams for roadmap, and evolving project governance
- 31 founding member organizations
  - 13 Premier Members, 10 General Members, 8 Associate members (academic and government institutions)
- 3 more members have joined since launch

CEPHALOCON BEIJING

- Inaugural Cephalocon APAC took place in March 2018
  - Beijing, China
  - 2 days, 4 tracks, 1000 attendees
  - Users, vendors, partners, developers
- 14 industry sponsors

CEPHALOCON BARCELONA

- Cephalocon Barcelona 2019
  - May 19-20, 2019
  - Barcelona, Spain
  - Similar format: 2 days, 4 tracks
- Co-located with KubeCon + CloudNativeCon
  - May 20-23, 2018
- CFP closed yesterday!
- Early-bird registration through Feb 15
  - Reduced hobbyist rate also available
- https://ceph.com/cephalocon/

THANK YOU

http://ceph.io/
sage@redhat.com
@liewegas