# ceph

## what's new in mimic and beyond

Kefu Chai

ceph

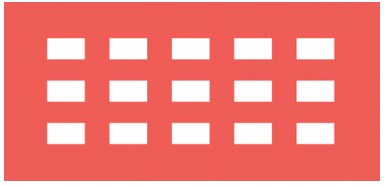
# Agenda

- Ceph
- what’s new in Mimic
- Dashboard v2
- Centralized config
- Nautilus

# Ceph

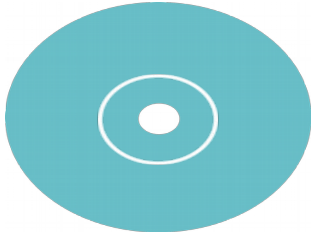
- Unified, distributed storage system
- Scalable
  - 10s to 1000s of OSDs (storage daemons)
- Hardware agnostic
  - HDD, SSD, whatever; no RAID required
  - IP network
- Fault tolerant
- Elastic
  - Dynamically, transparently migrate data on failure, expansion, contraction

# CEPH UNIFIED STORAGE



OBJECT STORAGE

- S3 & Swift
- Multi-tenant
- Geo-Replication
- Native API



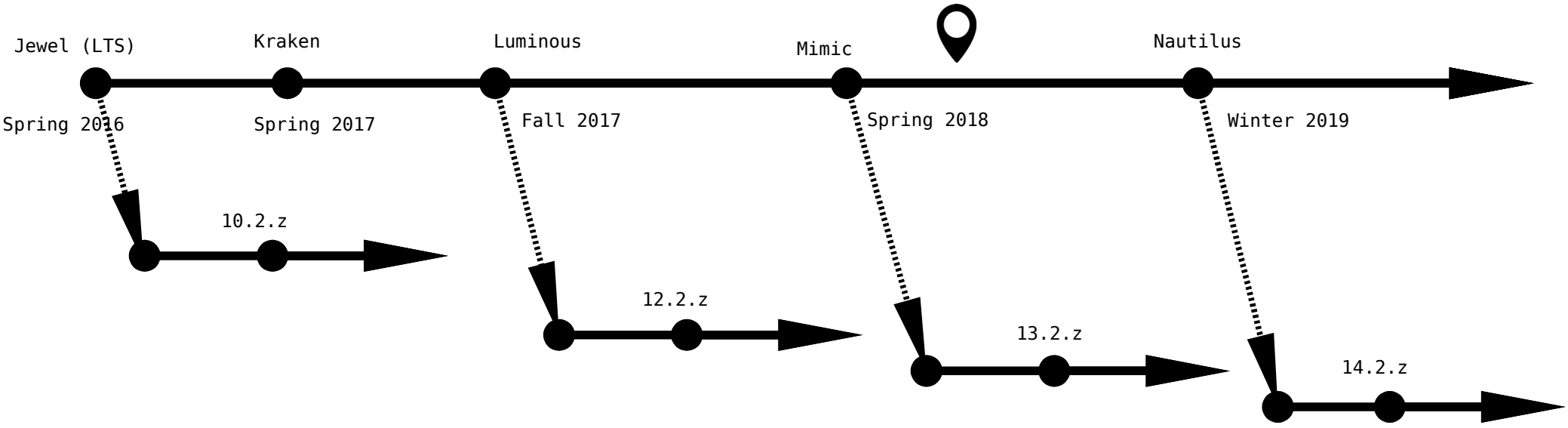
BLOCK STORAGE

- Snapshots
- Cloning
- OpenStack
- Linux Kernel
- iSCSI

FILE SYSTEM

- POSIX
- Linux Kernel
- CIFS/NFS
- Distributed Metadata

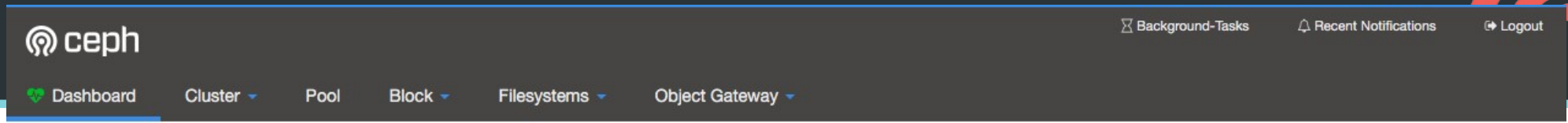
# Ceph Releases



# What's new in Mimic

- RADOS
  - ceph-mgr
    - Dashboard v2
  - ceph-mon
    - Automatic map pruning
  - ceph-osd
    - Async recovery
    - Allow client requests to preempt scrub
  - Centralized configuration management
  - Async interface in librados for use with Networking TS
- RBD (block)
  - supports clone of non-protected snapshot
- RGW (object)
  - Beast frontend
  - Cloud sync module
  - MFA support
  - AWS Bucket Policy
- CephFS (fs)
  - Quota support in kernel client 4.17 and higher
  - Many fixes for meta data balancer

# Dashboard v2




## Health

Overall status: HEALTH_OK

MONITORS
**3 (quorum 0, 1, 2)**

OSDS
**3 (3 up, 3 in)**

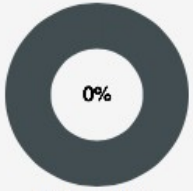
METADATA SERVERS
**1 active, 0 standby**

MANAGER DAEMONS
**active: x, 2 standbys**

## Usage

**22**

Objects



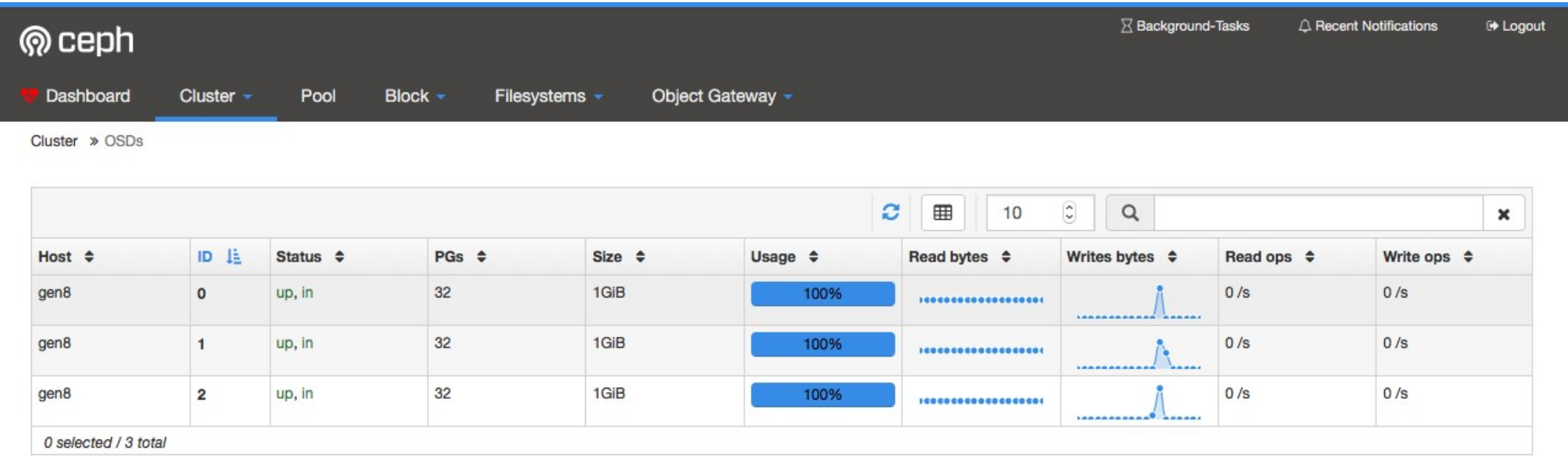

Raw capacity (99KiB used)

Usage by pool

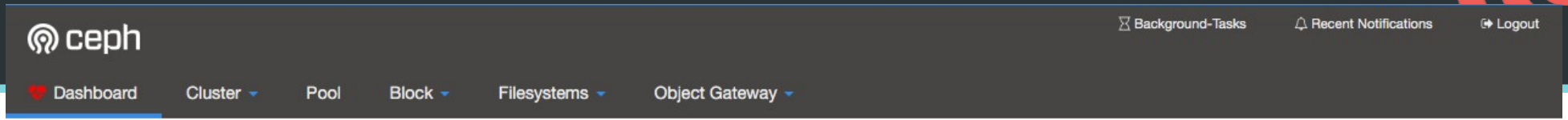
## Pools

| Name | PG status | Usage | Read | Write |
|---|---|---|---|---|
| cephfs_data_a | 8 active+clean | 0% | - - ops | - - ops |
| cephfs_metadata_a | 8 active+clean | 0% | - - ops | - - ops |

# Dashboard v2



# Dashboard v2




## Health

Overall status: HEALTH_ERR

- OSD_FULL: 3 full osd(s)
- POOL_FULL: 3 pool(s) full
- POOL_APP_NOT_ENABLED: application not enabled on 1 pool(s)

MONITORS
**3 (quorum 0, 1, 2)**

OSDS
**3 (3 up, 3 in)**

METADATA SERVERS
**1 active, 0 standby**

MANAGER DAEMONS
**active: x, 2 standbys**

## Usage

**470**

Objects

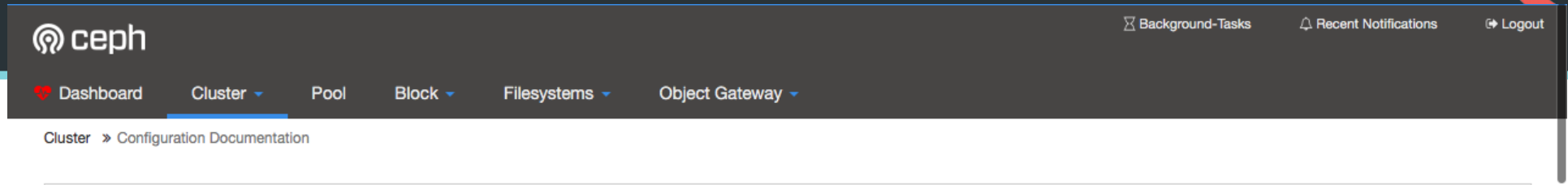

Raw capacity (3GiB used)

Usage by pool

## Pools

| Name | PG status | Usage | Read | Write |
|---|---|---|---|---|
| cephfs_data_a | 8 active+clean | NaN% | - - ops | - - ops |
| cephfs_metadata_a | 8 active+clean | 100% | - - ops | - - ops |
| scbench | 16 active+clean | 100% | - - ops | - - ops |

# Dashboard v2




Cluster » Configuration Documentation

Service: any Level: basic

| Name | Description | Type | Level | Default | Tags | Services | See_also | Max | Min |
|---|---|---|---|---|---|---|---|---|---|
| client_cache_size | soft maximum number of directory entries in client cache | int64_t | basic | 16384 | | mds_client | | | |
| cluster_addr | cluster-facing address to bind to | entity_addr_t | basic | - | network | osd | | | |
| err_to_graylog | send critical error log lines to remote graylog server | bool | basic | false | | | log_to_graylog<br>log_graylog_host<br>log_graylog_port | | |
| err_to_stderr | send critical error log lines to stderr | bool | basic | false true | | | | | |
| err_to_syslog | send critical error log lines to syslog facility | bool | basic | false | | | | | |
| fsid | cluster fsid (uuid) | uuid_d | basic | 00000000-0000-0000-0000-000000000000 | service | common | | | |
| host | local hostname<br>if blank, ceph assumes the short hostname (hostname -s) | std::string | basic | | network | common | | | |
| log_file | path to log file | std::string | basic | /var/log/ceph/$cluster-$name.log | | | log_to_stderr<br>err_to_stder | | |

# Config – new commands

```
$ ceph config -h
[...]
config assimilate-conf                  Assimilate options from a conf, and return a
                                         new, minimal conf file
config dump                             Show all configuration option(s)
config get <who> {<key>}                Show configuration option(s) for an entity
config help <key>                       Describe a configuration option
config log {<int>}                      Show recent history of config changes
config reset <int>                      Revert configuration to previous state
config rm <who> <name>                  Clear a configuration option for one or more
                                         entities
config set <who> <name> <value>         Set a configuration option for one or more
                                         entities
config show <who> {<key>}               Show running configuration
config show-with-defaults <who>         Show running configuration (including
                                         compiled-in defaults)
config-key dump {<key>}                 dump keys and values (with optional prefix)
config-key exists <key>                 check for <key>'s existence
config-key get <key>                    get <key>
config-key ls                           list keys
config-key rm <key>                     rm <key>
config-key set <key> {<val>}            set <key> to value <val>
```

# Config

```
$ ceph config dump
WHO     MASK LEVEL    OPTION                          VALUE          RO
global       advanced mon_pg_warn_min_per_osd         3
global       advanced osd_pool_default_min_size       1
global       advanced osd_pool_default_size           3
  mon        advanced mon_allow_pool_delete           true
  mon        advanced mon_data_avail_crit             1
[...]
  mgr        unknown  mgr/restful/x/server_port       42976           *
  mgr        unknown  mgr/restful/y/server_port       44976           *
  mgr        unknown  mgr/restful/z/server_port       46976           *
  osd        advanced osd_copyfrom_max_chunk          524288
  osd        dev      osd_debug_misdirected_ops       true
  osd        dev      osd_debug_op_order              true
  osd        advanced osd_scrub_load_threshold        2000.000000
  mds        dev      mds_debug_auth_pins             true
  mds        dev      mds_debug_frag                  true
  mds        dev      mds_debug_subtrees              true
```

# Config

```
$ ceph config get 'osd.*' debug_ms
0/5
$ ceph config set osd debug_ms 1
$ ceph config dump
WHO    MASK LEVEL    OPTION                                VALUE RO
global      advanced mon_pg_warn_min_per_osd               3
global      advanced osd_pool_default_min_size             1
[...]
osd         advanced debug_ms                              1
$ ceph config get 'osd.*' debug_ms
1/1
$ ceph config get osd.0 debug_ms
1/1
```

# Config - overrides

```
$ ceph config set osd/class:hdd debug_ms 2
$ ceph config get 'osd.*'
WHO    MASK      LEVEL    OPTION                     VALUE       RO
osd    class:hdd advanced debug_ms                    2/2
global           advanced mon_pg_warn_min_per_osd    3
osd              advanced osd_copyfrom_max_chunk     524288
osd              dev      osd_debug_misdirected_ops true
[...]


$ ceph config rm osd/class:hdd debug_ms
$ ceph config get osd.0
WHO    MASK LEVEL    OPTION                     VALUE       RO
osd         advanced debug_ms                   1/1
osd.0       advanced debug_osd                  10/10
global      advanced mon_pg_warn_min_per_osd    3
osd         advanced osd_copyfrom_max_chunk     524288
osd         dev      osd_debug_misdirected_ops true
[...]
```

# Config – overrides (cont.)

```
$ ceph config set osd.0 debug_osd 10
$ ceph config get osd.0
WHO    MASK       LEVEL    OPTION                          VALUE        RO
osd    class:hdd  advanced debug_ms                        3/3
osd.0             advanced debug_osd                       10/10
global            advanced mon_pg_warn_min_per_osd        3
osd               advanced osd_copyfrom_max_chunk         524288
osd               dev      osd_debug_misdirected_ops      true
[...]


$ ceph daemon osd.0 config set debug_osd 10
{
    "success": ""
}
$ ceph config show osd.0
NAME                          VALUE         SOURCE    OVERRIDES  IGNORES
[...]
bluestore_block_wal_size      1048576000    file
bluestore_fsck_on_mount       true          file
chdir                                       file
debug_ms                      1/1           mon
```

# Config – more typed settings

```
$ ceph config set osd.10 osd_scrub_max_preemptions -1
Error EINVAL: error parsing value: strict_sistrtoll: value should not be negative
$ ceph config set osd.10 osd_scrub_max_preemptions 1k
Error EINVAL: error parsing value: strict_si_cast: unit prefix not recognized
$ ceph config help osd_scrub_max_preemptions
osd_scrub_max_preemptions - Set the maximum number of times we will preempt a deep
scrub due to a client operation before blocking client IO to complete the scrub
  (uint64_t, advanced)
  Default: 5
  Can update at runtime: true
$ ceph config set osd.10 osd_scrub_max_preemptions 1K

$ ceph config help mon_op_complaint_time
mon_op_complaint_time - time after which to consider a monitor operation blocked
after no updates
  (secs, advanced)
  Default: 30
  Can update at runtime: true
$ ceph config set mon mon_op_complaint_time 13days
$ ceph config get mon.a mon_op_complaint_time
1123200
```

# Config – log

```
$ ceph config log
--- 31 --- 2018-06-17 11:18:40.988109 ---
+ osd.10/osd_scrub_max_preemptions = 1000
--- 30 --- 2018-06-17 11:09:04.993020 ---
- osd/class:hdd/debug_ms = 3/3
--- 29 --- 2018-06-17 11:08:44.478144 ---
- osd/class:ssd/debug_ms = 2/2
--- 28 --- 2018-06-17 11:04:20.932011 ---
+ osd.0/debug_osd = 10/10
--- 27 --- 2018-06-17 11:01:43.019134 ---
+ osd/class:hdd/debug_ms = 3/3
[...]
$ ceph config reset 30
$ ceph config get osd.10 osd_scrub_max_preemptions
5
```

# Config – migrating from old configs

```
$ cat /etc/ceph/ceph.conf
[global]
mon host = foo.ceph.com
[osd.1]
debug_osd = 0/0
[mds.a]
mds invalid option = this option does not exist

$ ceph config assimilate-conf -i /etc/ceph/ceph.conf -o ceph.conf.new
[global]
        mon_host = foo.ceph.com

[mds.a]
        mds_invalid_option = this option does not exist

$ ceph config get osd.1
WHO    MASK LEVEL    OPTION           VALUE       RO
osd.1       advanced debug_osd        0/0

$ cat ceph.conf.new
$ mv ceph.conf.new /etc/ceph/ceph.conf
```

# Config

```
$ ceph config set osd.10 osd_scrub_max_preemptions -1
Error EINVAL: error parsing value: strict_sistrtoll: value should not be negative
$ ceph config set osd.10 osd_scrub_max_preemptions 1k
Error EINVAL: error parsing value: strict_si_cast: unit prefix not recognized
$ ceph config help osd_scrub_max_preemptions
osd_scrub_max_preemptions - Set the maximum number of times we will preempt a deep
scrub due to a client operation before blocking client IO to complete the scrub
  (uint64_t, advanced)
  Default: 5
  Can update at runtime: true
$ ceph config set osd.10 osd_scrub_max_preemptions 1K

$ ceph config help mon_op_complaint_time
mon_op_complaint_time - time after which to consider a monitor operation blocked
after no updates
  (secs, advanced)
  Default: 30
  Can update at runtime: true
$ ceph config set mon mon_op_complaint_time 13days
$ ceph config get mon.a mon_op_complaint_time
1123200
```

# Nautilus and beyond

- RADOS
  - Better QoS (dmClock)
  - On the wire encryption
  - Kerberos authentication
  - Clay codes
  - PG merge
  - Mgr
    - Captured crash report
    - Disk failure prediction
    - Orchestrator Python interfaces
  - Fully async OSD
- RBD (block)
  - Namespace support
- CephFS (FS)
  - Multiple independent CephFS filesystem

# THANK YOU!

Kefu Chai

kefu@
#ceph-devel

kefu@redhat.com