THE FUTURE OF STORAGE

SAGE WEIL – RED HAT

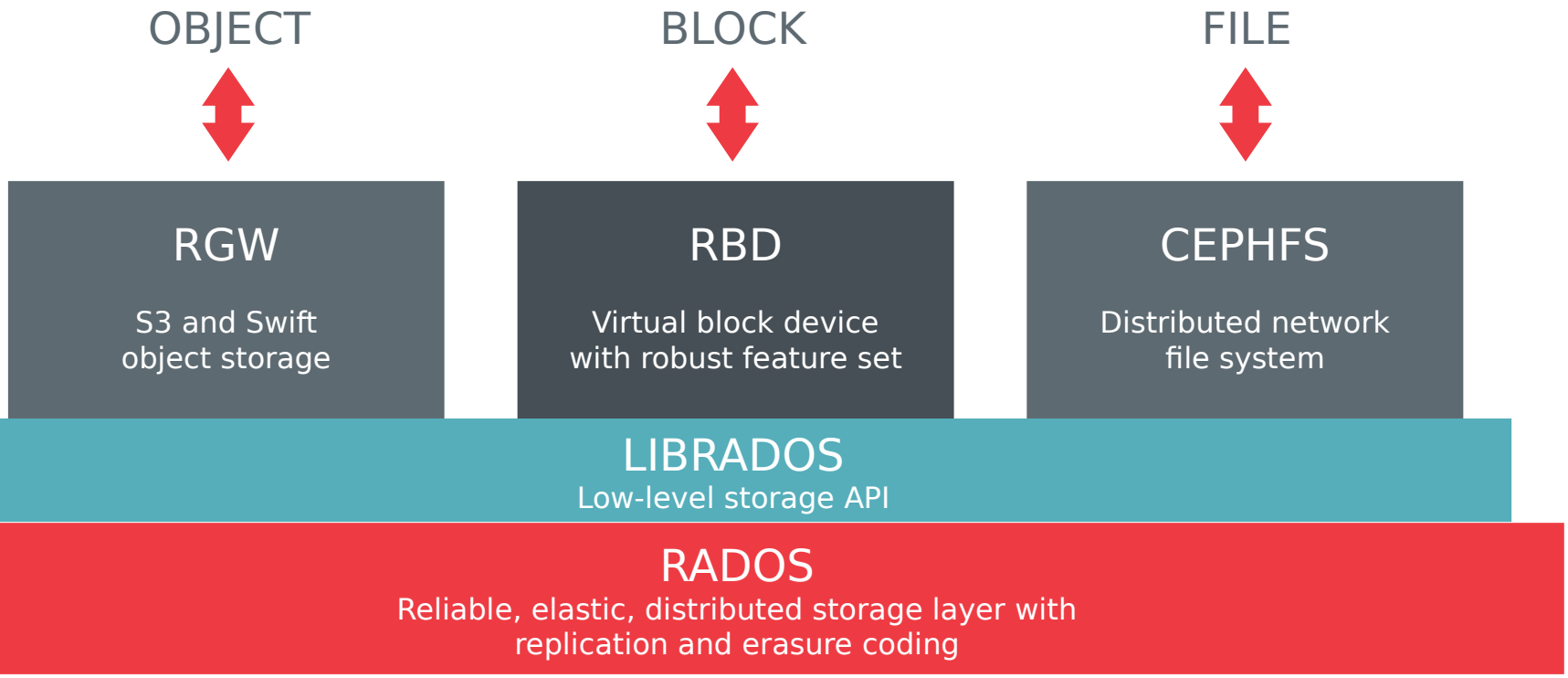
2018.03.22

WELCOME

THANK YOU

WHAT IS CEPH?

UNIFIED STORAGE SOFTWARE



CEPH IS FREE AND OPEN SOFTWARE

- Free software is a better development model
  - efficient use of resources, effort
  - enables free integration of complementary projects
- Free software is better for the user
  - commodity hardware components
  - flexibility
  - freedom from vendor lock-in
- Free software is better for the world

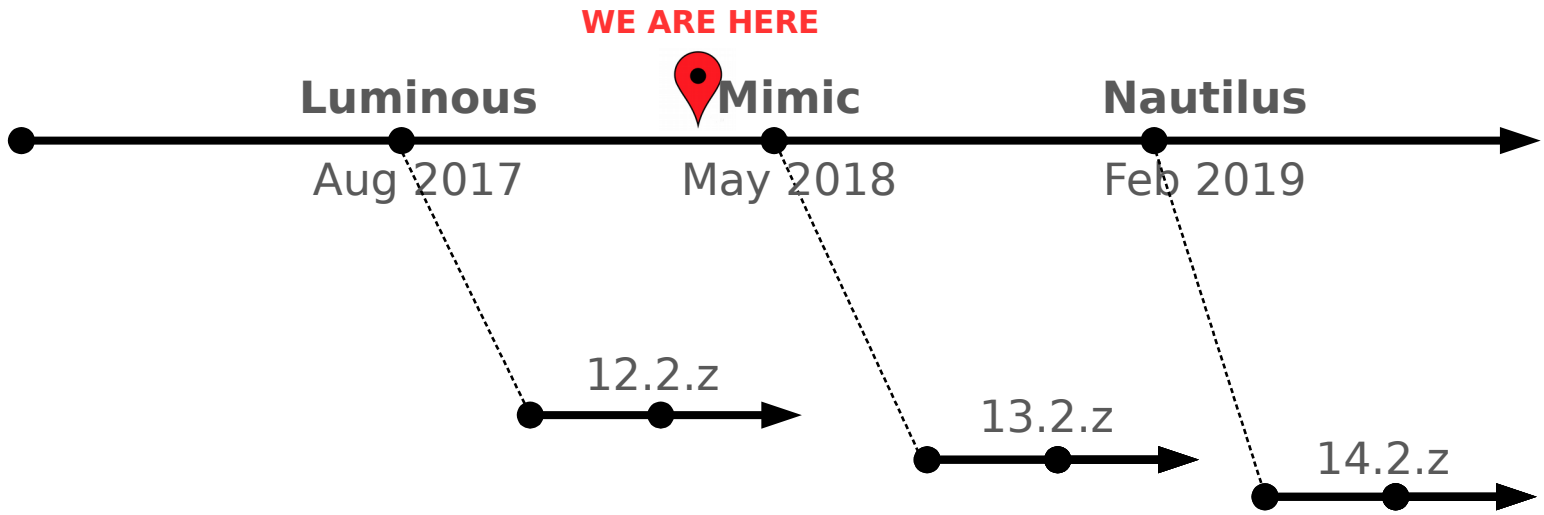
CEPH IS RELIABLE

- No single point of failure
- Reliable, durable storage service out of unreliable components
- Replication and erasure coding
- Favor consistency and correctness over performance

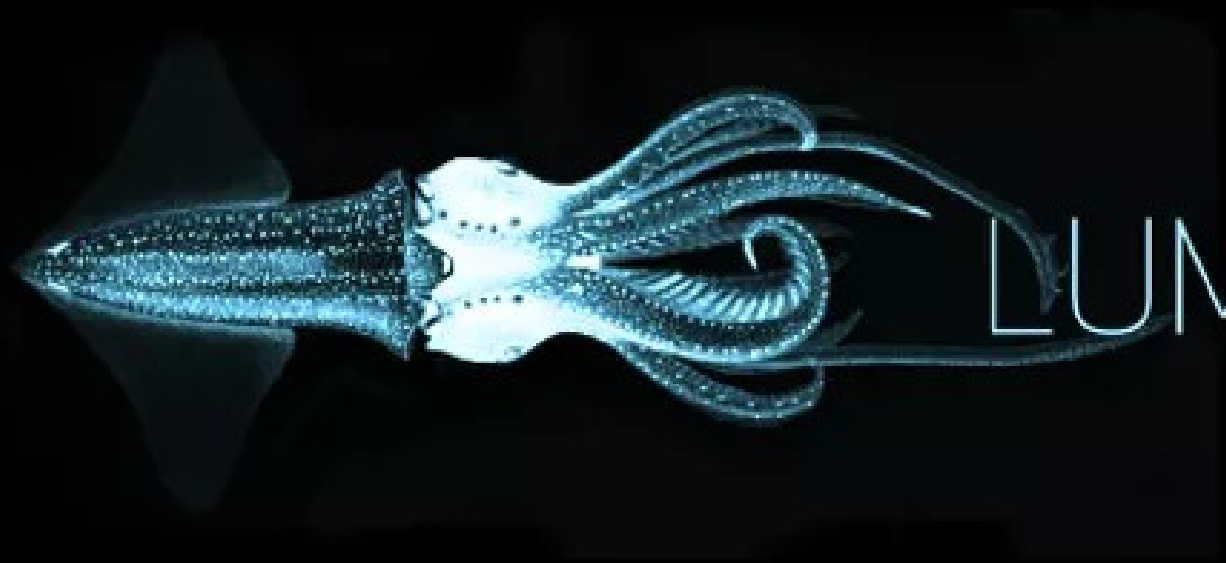
CEPH IS SCALABLE

- Designed for supercomputing workloads
- Storage needs grow over time
- Ceph is an elastic architecture
  - Add or remove storage hardware while cluster is online
- Online, rolling software upgrades
- Single-cluster scaling within a data center/region
- Multi-cluster federation of services across regions/geos

CEPH RELEASES



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



ceph

LUMINOUS

Multiple MDS for CephFS

RGW search

RGW NFS

PG balancing

BlueStore

iSCSI

Erasure code for RBD and CephFS

ceph-mgr

Better scalability

LUMINOUS

MIMIC

MIMIC HIGHLIGHTS

- RADOS
  - PG merging and autoscaling
  - mon-based config management
  - experimental quality-of-service

  Josh Durgin and Greg Farnum

- RBD (block)
  - image groups
  - deep copy of images
  - live migration

  Jason Dillaman

- RGW (object)
  - sync to cloud (S3)
  - new frontend (performance)

  Matt Benjamin and Orit Wasserman

- CephFS (file)
  - stable snapshots
  - kernel client quotas

  Patrick Donnelly

THE FUTURE

FOUR CHALLENGES AND OPPORTUNITI

- Performance
- New platforms
- Ease of use
- Scaling

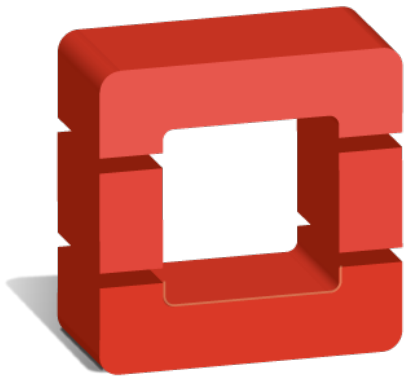
PERFORMANCE

CHANGING HARDWARE LANDSCAPE

- BlueStore significantly moves us forward
- HDDs relegated to streaming and archival workloads
- NVMe SSDs everywhere else
- Persistent memory is coming*
- CPU is the new bottleneck
- Significant refactoring of OSD is in progress
  - *Seastar, DPDK, SPDK*

NEW PLATFORMS

OPENSTACK PUSHED CEPH MAINSTREA



openstack™

ceph

CONTAINER ORCHESTRATION

- container is the new package
- Kubernetes is the new container platform of choice
- scale-out platforms need scale-out storage
- container orchestration is an enabler for hyperconverged infrastructure
- Rook
  - Ceph operator for Kubernetes
  - manages deployment lifecycle of a Ceph cluster

  Bassam Tabbara

EMERGING WORKLOADS

- Public cloud file as a service with CephFS (via OpenStack Manila)
- HPC with CephFS
- Data lake
  - object storage archives
  - big data analytics (Hadoop and Spark)
- Artificial intelligence and machine learning (via Spark)
- Imagery archives (medical, satellite)
- Scientific computing

EASE OF USE

CEPH MUST BE EASY

- Ceph must be self-managing in order to scale
- Must change Ceph's reputation for being "hard"
- Luminous brought huge list of improvements
- More coming in Mimic
  - pg_num, PG merging, and PG autotuning
  - centralized configuration management

NEW CEPH DASHBOARD

- Mimic will bring a refreshed/rebooted dashboard UI
  - SUSE’s openATTIC merged upstream into ceph-mgr
  - combined effort moving forward by community
- Included in core Ceph, works out of the box
- management and monitoring
  - integration with deployment tools for, e.g., expansion, hardware maintenance

SCALE

TESTING AT SCALE

- Ceph is designed to enable big clusters
- Developers cannot afford that much hardware
- CERN “big bang” tests
  - 2015 ~7,000 OSDs with Ceph Hammer
  - 2016 ~7,000 OSDs with Ceph Jewel
  - 2017 ~10,000 OSDs with Ceph Luminous

- Key challenge is to test at scale with real workloads

SCALING BEYOND ONE CLUSTER

- Live in a multi-cloud and hybrid cloud world
  - on premise deployments, public cloud, managed cloud
- Multi-regional redundancy and disaster recovery
- RGW object storage
  - Now: multi-cluster federation and replication
  - Future: replication to public cloud object storage
- RBD block storage
  - Now: multi-cluster asynchronous replication
  - Future: better orchestration tools to migrate workloads *and* storage across clouds
- CephFS file storage
  - Future: multi-datacenter replication

GET INVOLVED

- Mailing list and IRC
  - http://ceph.com/IRC
- Github
  - https://github.com/ceph/
- Ceph Developer Monthly
  - first Weds of every month
  - video conference (Bluejeans)
  - alternating APAC- and EMEA-friendly times
- Cephalocon!
- Ceph Days
  - http://ceph.com/cephdays/
- Meetups
  - http://ceph.com/meetups
- Ceph Tech Talks
  - http://ceph.com/ceph-tech-talks/

THANK YOU

THANK YOU