Introduction to CEPH

CEPH- What?

Distributed storage platform which implements object storage without a single point of failure, scalable to exabyte level and freely available

CEPH - Why?

Traditional storage systems are not equipped for the amount of data generated nowadays . Software defined storage is the best alternative.

ceph

CEPH - What makes Ceph unique?

Design Principles :

- Scalable
- No Single point of failure
- Software based
- Self Managing

Philosophy :

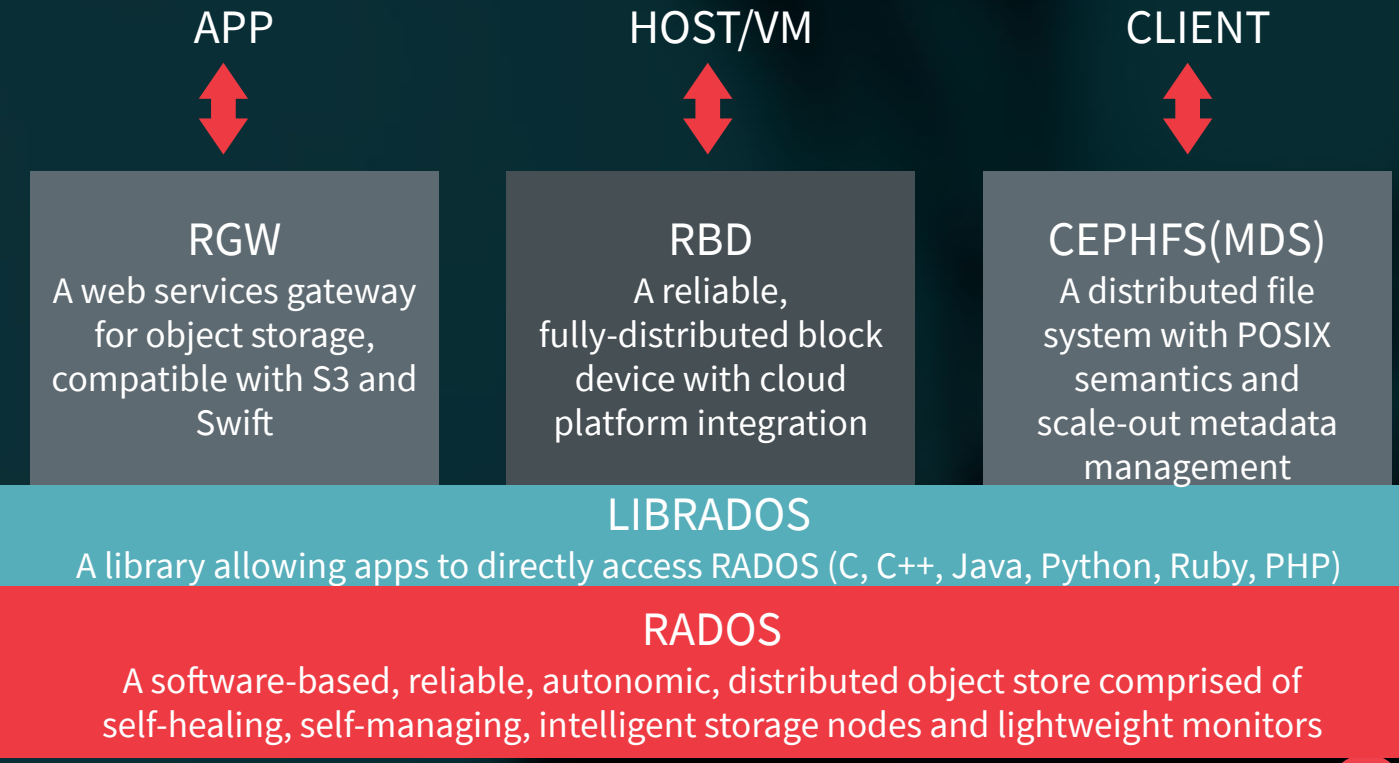
- Open source
- Community based

ceph

CEPH ARCHITECTURE

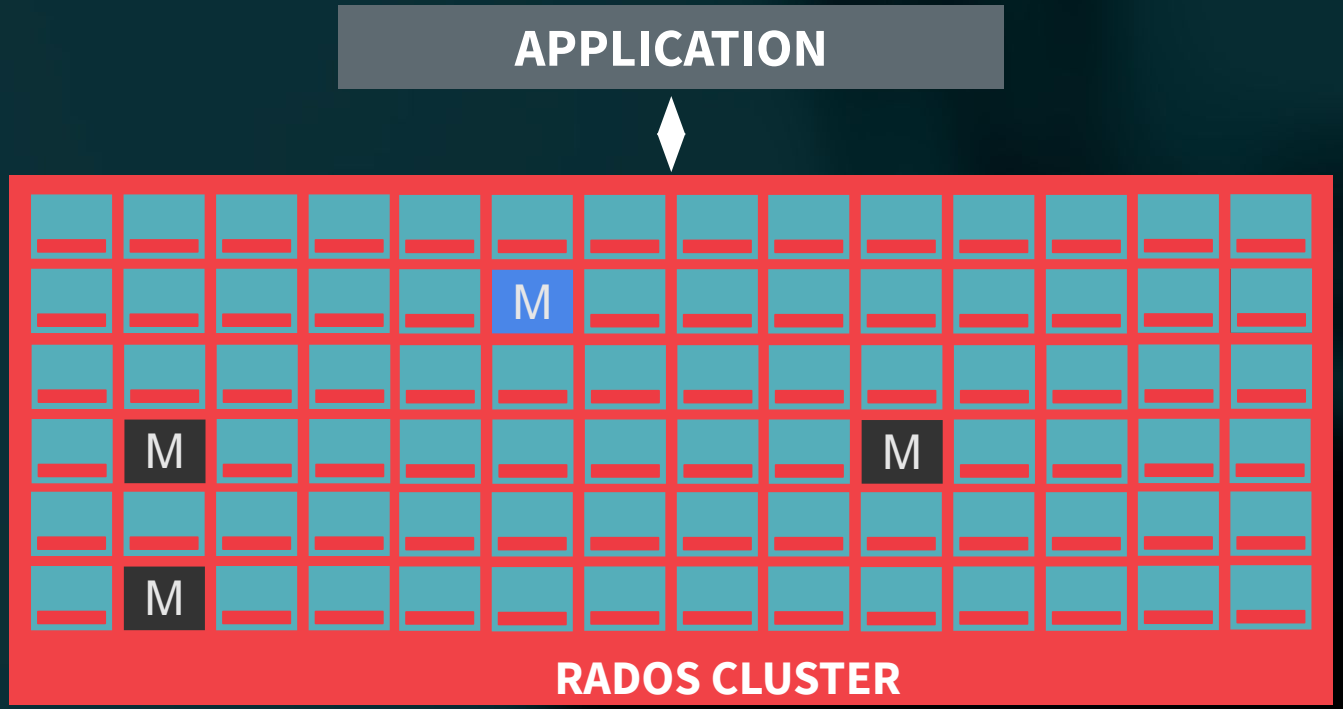
ceph

ARCHITECTURAL COMPONENTS



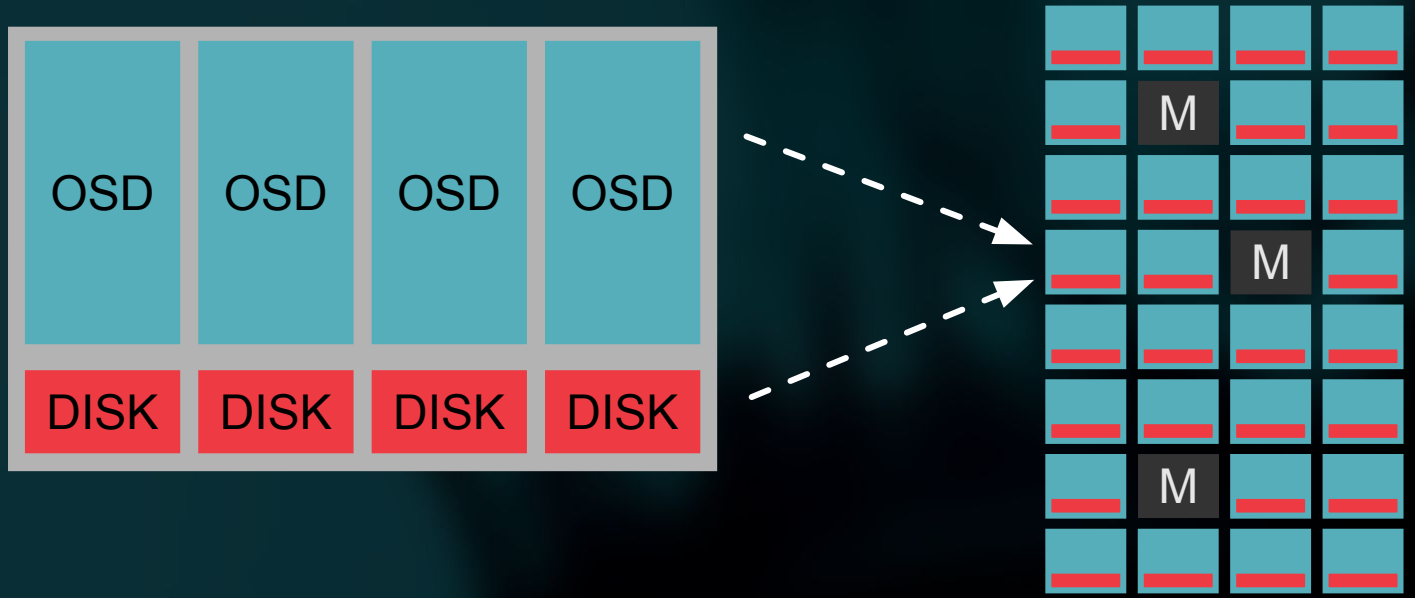
ceph

RADOS CLUSTER



ceph

OBJECT STORAGE DAEMONS



ceph

RADOS COMPONENTS

OSDs:

- 10s to 10000s in a cluster
- One per disk (or one per SSD, RAID group…)
- Serve stored objects to clients
- Intelligently peer for replication & recovery

Monitors:

- Maintain cluster membership and state
- Keeps the state of all MAPS.
- Small, odd number
- Do not serve data

ceph

Ceph Manager

Managers:

- Tightly coupled with Monitor
- Manage cluster “logistic” PG and maps
- Provide additional monitoring and interfaces to external monitoring and management systems
- Pluggable python interface to develop modules
- Some modules:
  - Balancer
  - Dashboard
  - RESTful
  - Prometheus

ceph

Responsibilities of RADOS Cluster :

- Write and read data
- Ensure durability by replicating or erasure coding data
- Monitor and report on cluster health—also called 'heartbeating'
- Redistribute data dynamically—also called 'backfilling'
- Ensure data integrity
- Recover from failures

ceph

Some Concepts of RADOS Cluster :

- Pools
  - Replicated
  - EC
- PG’s
- Authentication
- CRUSH

ceph

OBJECT PLACEMENT WITH CRUSH

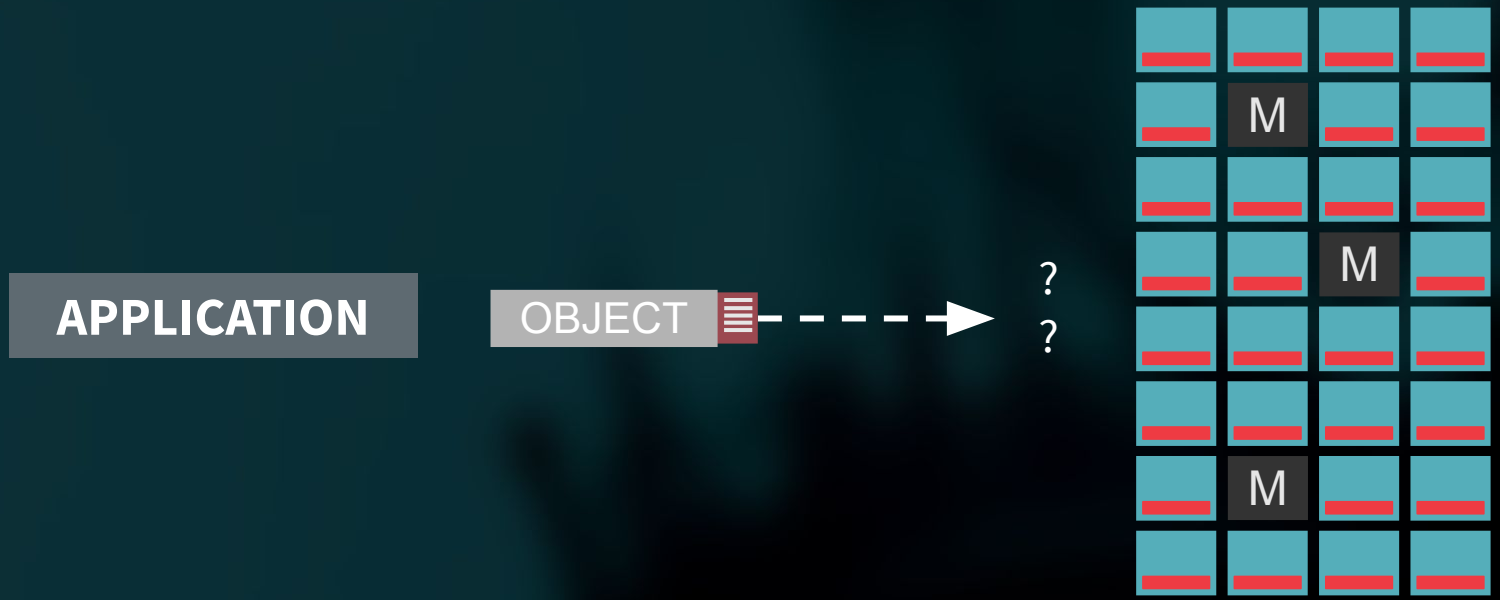
ceph

Client Communication with Ceph Storage Cluster

- Ceph Configuration File
- Pool Name
- User name and the path to the secret key

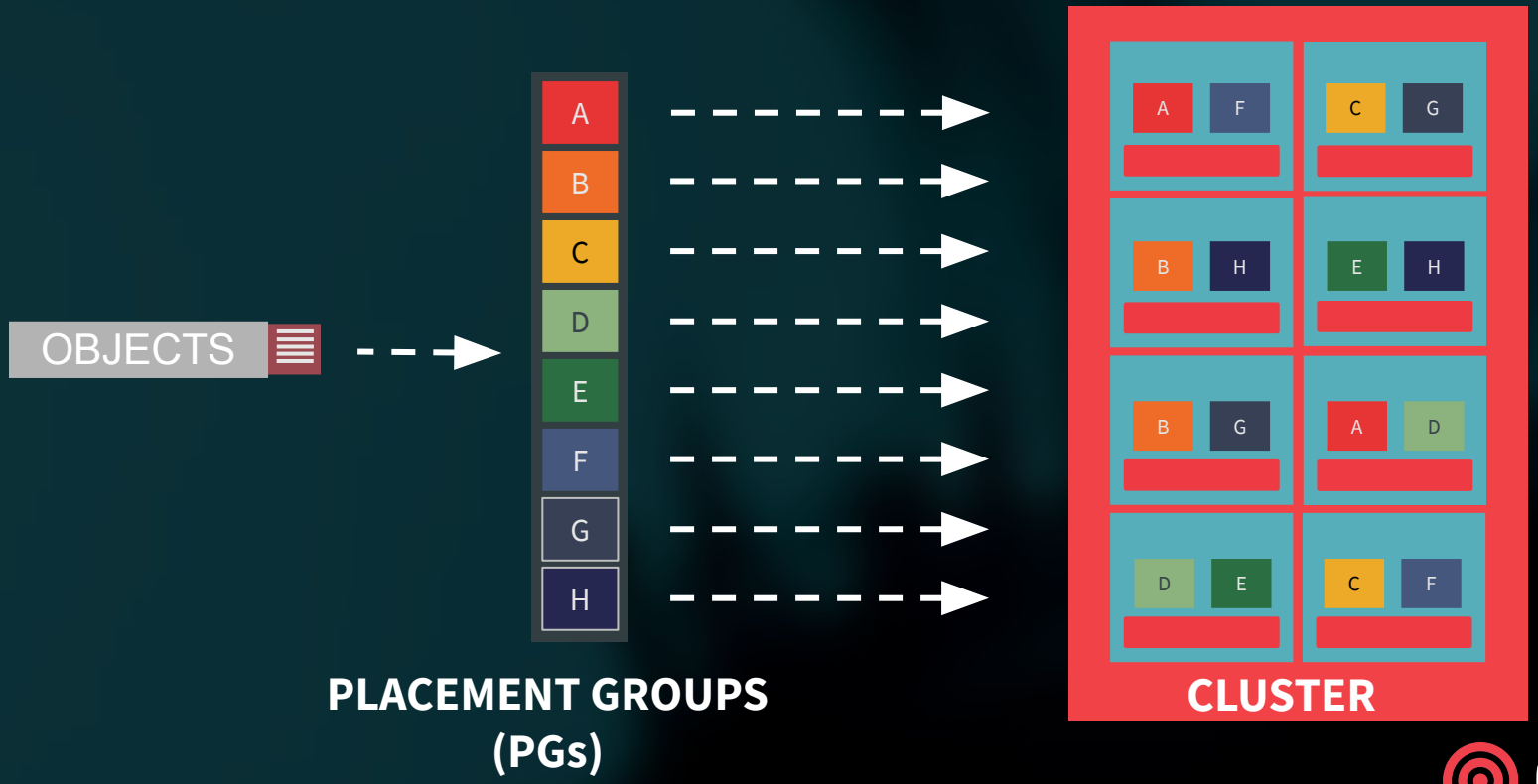
ceph

WHERE DO OBJECTS LIVE?



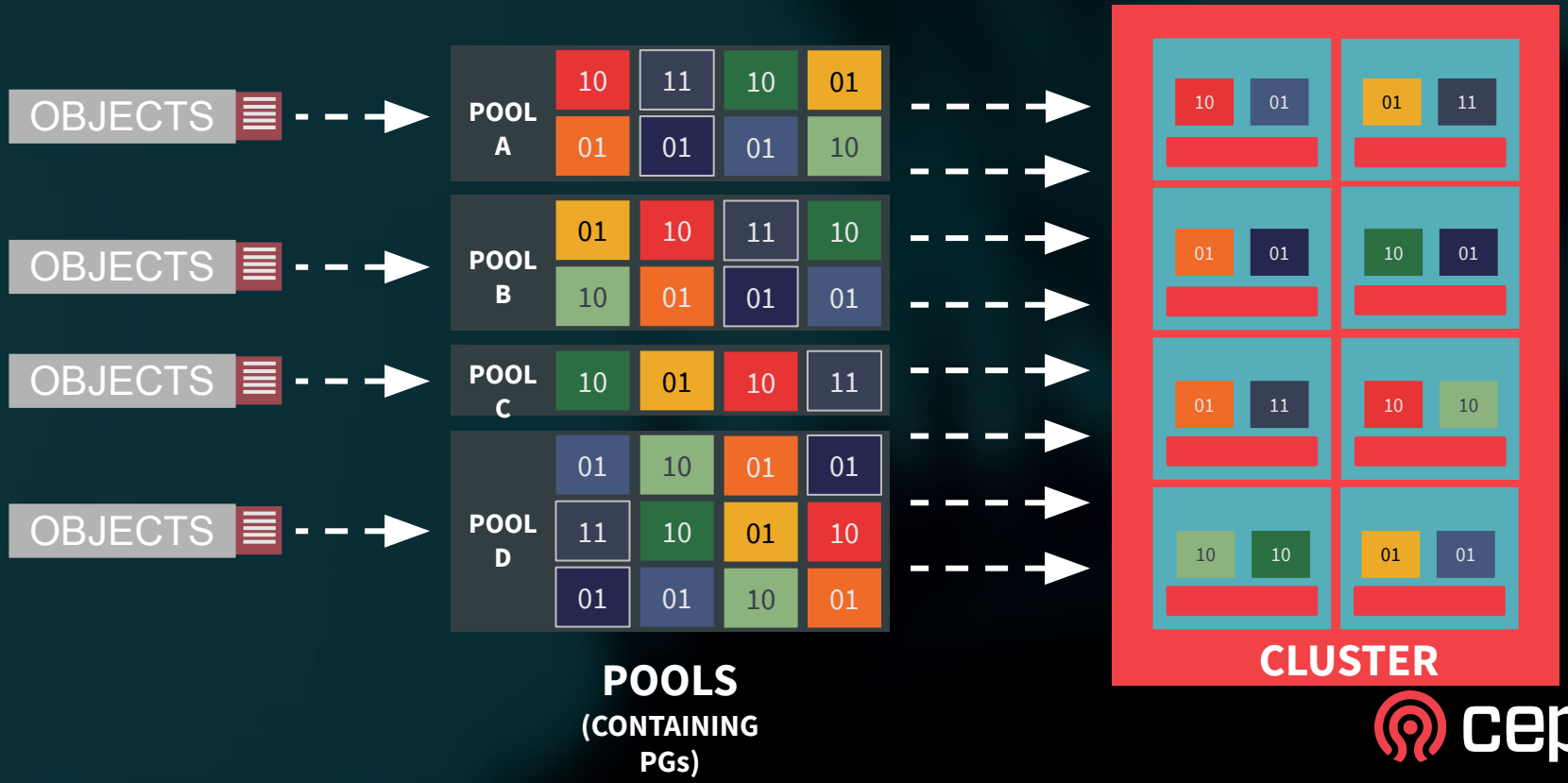
ceph

CRUSH: placement algorithm



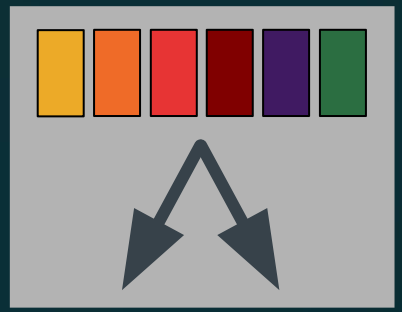
ceph

DATA IS ORGANIZED INTO POOLS



ceph

CRUSH: DYNAMIC DATA PLACEMENT



CRUSH (Controlled Replication Under Scalable Hashing)

- Pseudo-random placement algorithm
  - Fast calculation, no lookup
  - Repeatable, deterministic
- Statistically uniform distribution
- Stable mapping
  - Limited data migration on change
- Rule-based configuration
  - Infrastructure topology aware
  - Adjustable replication
  - Weighting

ceph

ceph