CEPH: IS IT AN INTERESTING SOLUTION IN THE LONG TERM FUTURE?

Giacinto Donvito
INFN-BARI

CEPH: CONCEPT AND ARCHITECTURE

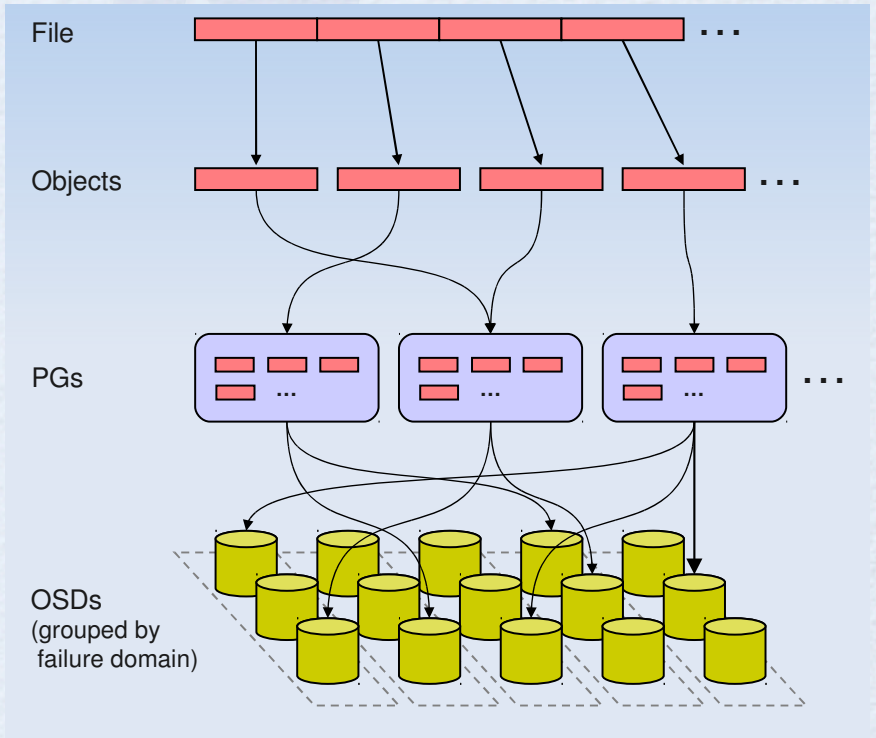
- An object based parallel file-system
- Open source project (LGPL licensed)
- Written in C++ and C
- kernel level
- Posix compliant
- No SPOF
  - Both data and metadata could be replicated dynamically
- Configuration is config file based
- Flexible striping strategies and object sizes
  - Could be configured “per file”

CEPH: CONCEPT AND ARCHITECTURE

- Key goals:
  - Scale up to 10'000 storage servers
  - Petabytes of data
  - TB/sec aggregate throughput
  - Billions of files organized in one to thousands of files per directory
  - File sizes that range from bytes to terabytes
  - Metadata access times in µsecs
- Object and metadata are replicated using a policy based algorithm that could be easily configured by sysadmin
- SAN (shared) disk is not needed to achieve HA

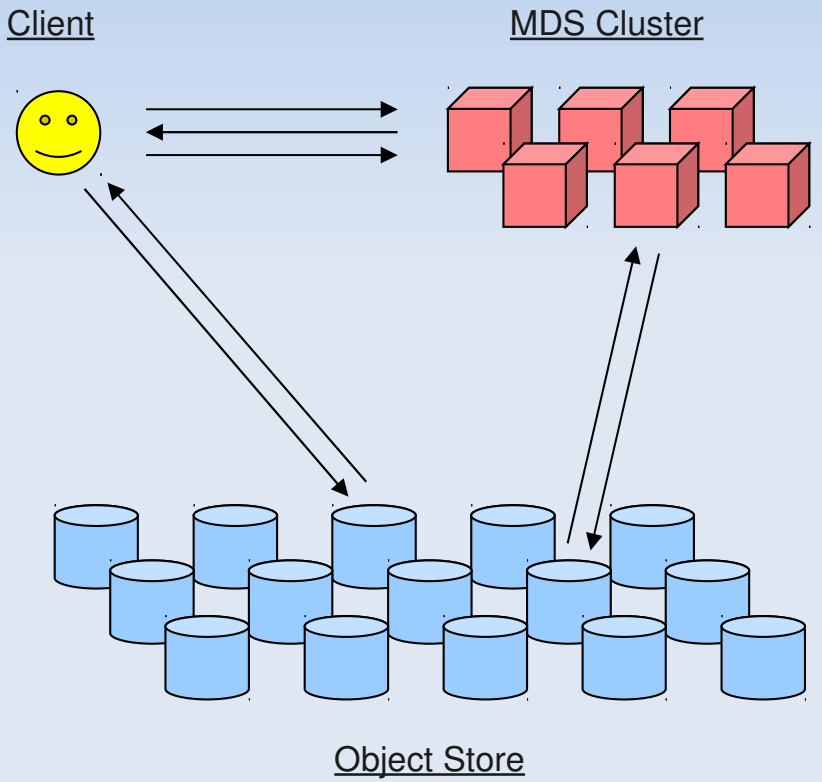
CEPH: CONCEPT AND ARCHITECTURE

- Data Placement is realized by means of “hash functions”:
  - Location of data is calculated => no lookup tables
    - this means: unstable mapping and adding disk servers means reshuffling
    - “Rules” driven by replica: “three replica should be in different cabinet”
    - three concept could be used:
      - Disk, Server, Rack



CEPH: CONCEPT AND ARCHITECTURE

- fd=open("/foo/bar", O_RDONLY)
  - Client: requests open from MDS
  - MDS: reads directory /foo from object store
  - MDS: issues capability for file content
- read(fd, buf, 1024)
  - Client: reads data from object store
- close(fd)
  - Client: relinquishes capability to MDS

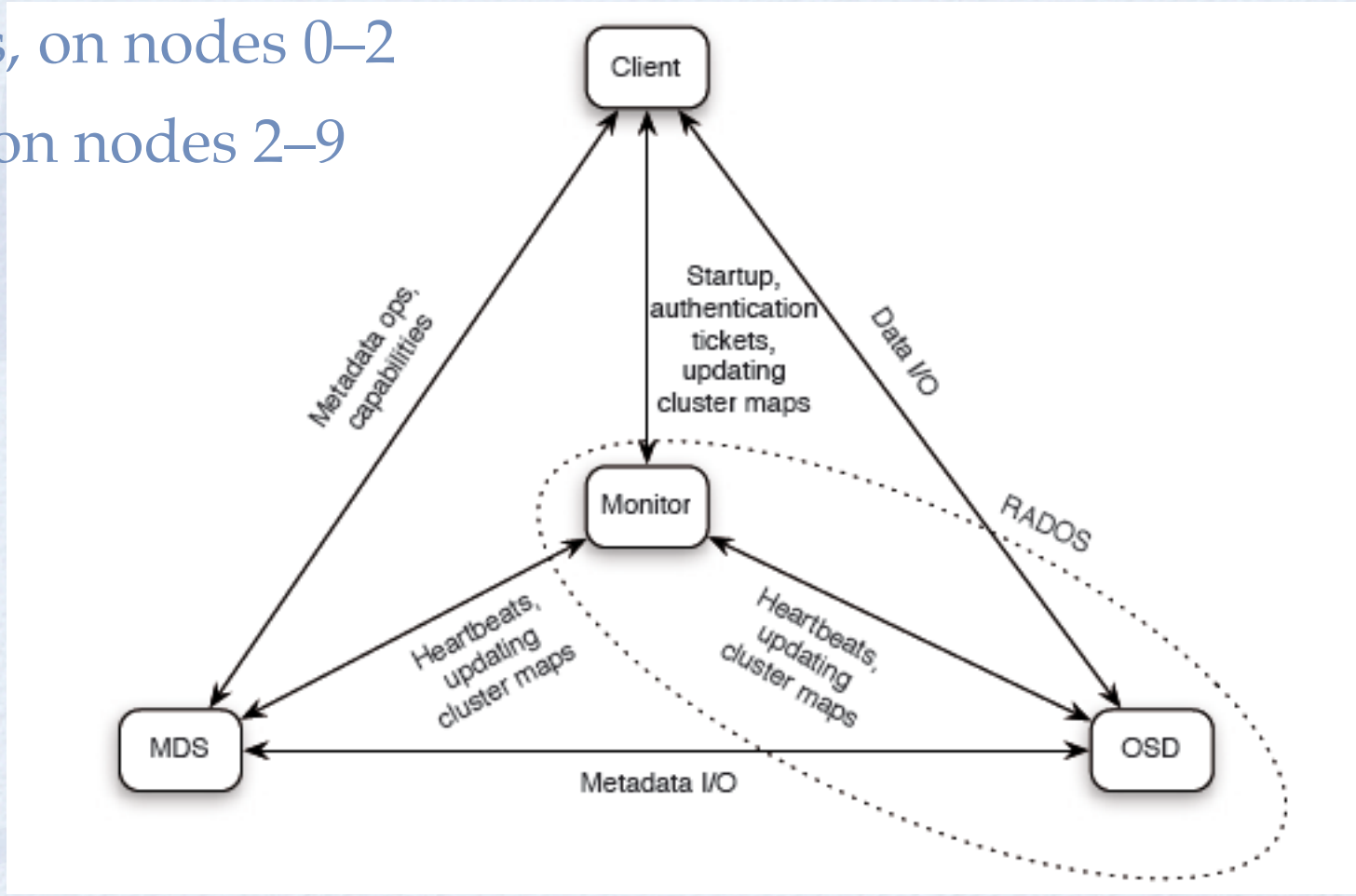


- MDS out of I/O path
- Object locations are well known–calculated from object name

CEPH: CONCEPT AND ARCHITECTURE

- For example having 10 nodes, we can configure:
  - Three monitors, on nodes 0–2
  - Three MDSes, on nodes 0–2
  - Eight OSDs, on nodes 2–9

NO SPOF

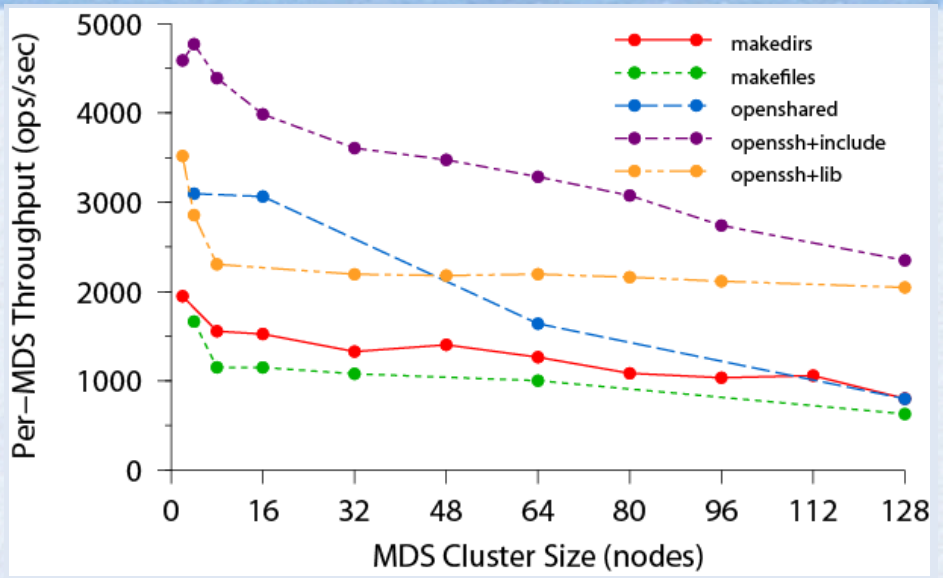


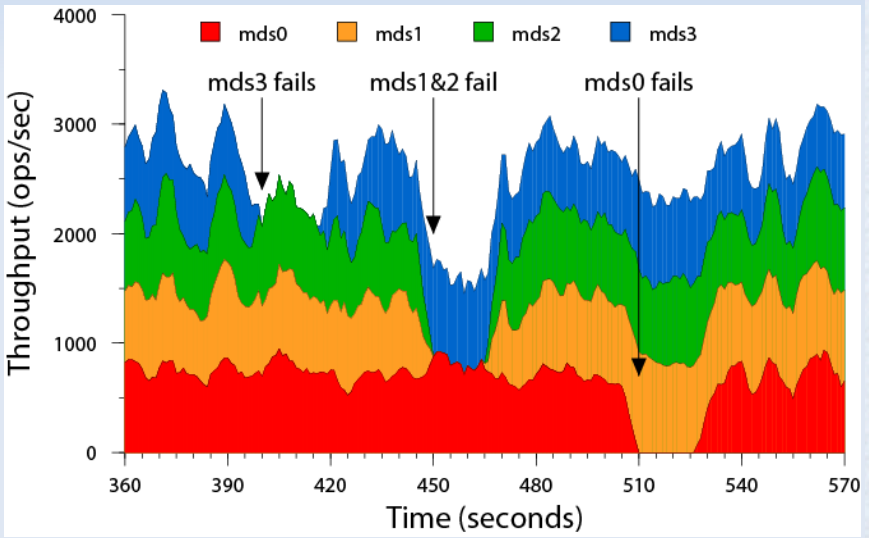
CEPH: CONCEPT AND ARCHITECTURE

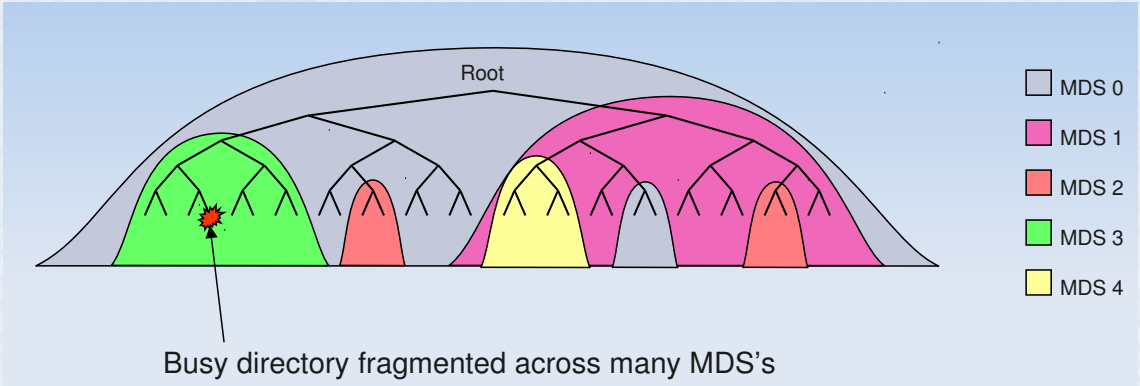
- Intelligent server: replicate data, migrate object, detect node failures
  - this could happen because everyone know where object belongs
- inodes are stored together with the directory object: you can load complete directory and inodes with a single I/O ("find" or "du" are greatly faster)
- It is easy to build a cluster of metadata servers (MDS)
  - Than it is scalable and adaptive
    - The work is moved from busy servers to idle ones

CEPH: CONCEPT AND ARCHITECTURE

- Up to 128 MDS nodes and 250kops/s
- I/O rates of potentially many TB/s
- File system containing many petabytes of storage





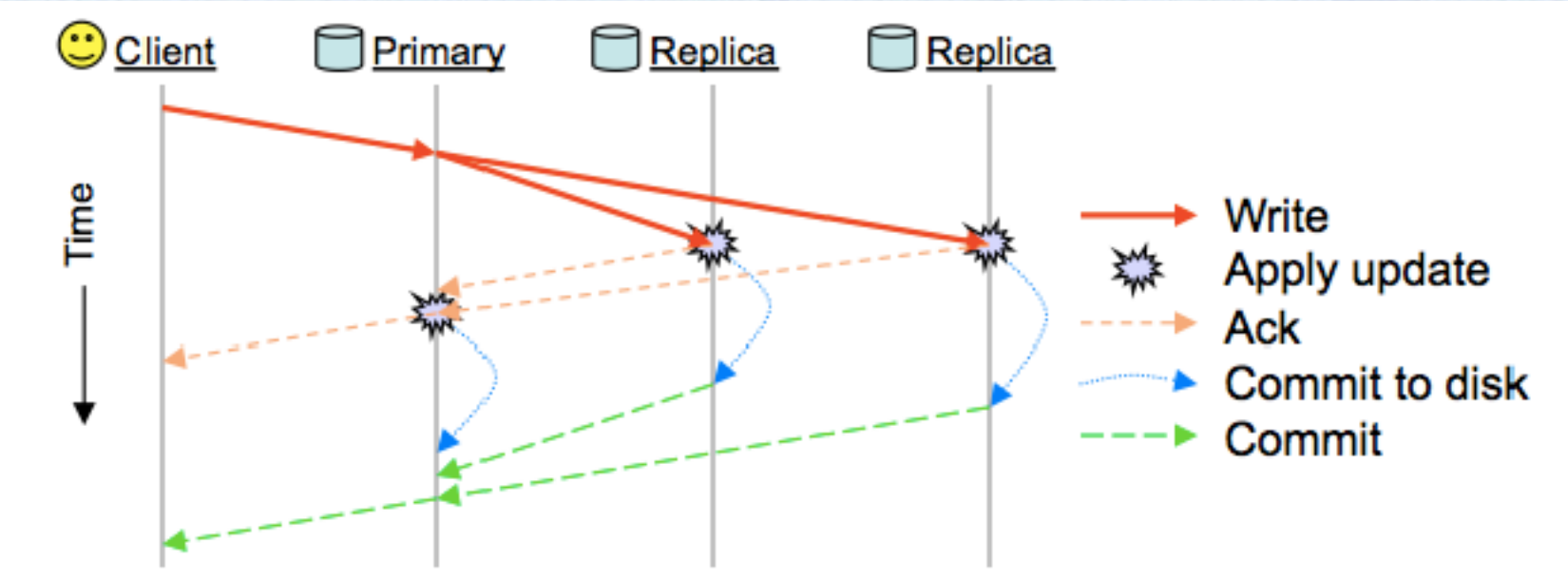


CEPH: CONCEPT AND ARCHITECTURE

- Subtree based usage accounting (half the work of a quota system)
- Near-posix, strong consistency
- Support snapshots
- kernel > 2.6.34 is required on client side or is there a FUSE client with the kernel > 2.6.35 the server is already built-in

```
$ ls -al
drwx------ 1 root root 5438384 Oct 20 14:51 ./
drwx------ 1 root root 5438387 Oct 20 14:51 ../
drwxr-xr-x 1 root root 2342034 Apr 20  2009 ghostscript/
drwxr-xr-x 1 root root  276961 Apr 20  2009 libthai/
drwx------ 1 root root 2817666 Oct 20 14:51 python-support/
drwxr-xr-x 1 root root    1723 Apr 20  2009 readline/
```

CEPH: WRITE SEMANTICS



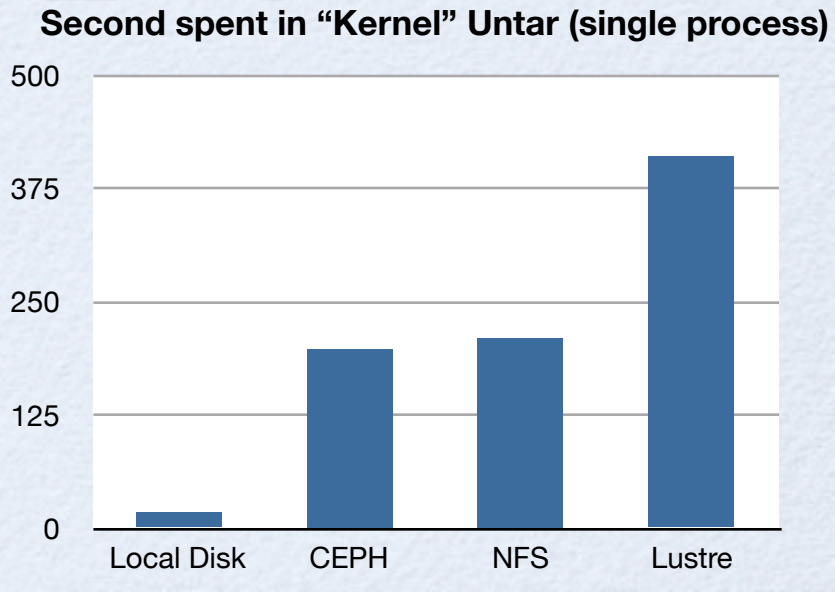
- By default, OSDs use Btrfs as their local file system (but ext3 works too). Data is written asynchronously using copy-on-write, so that unsuccessful write operations can be fully rolled back.

CEPH: FIRST TEST AND FEEDBACK

- Using standard Ubuntu 10.10
  - It is already compiled and available on default kernel (2.6.35)
  - Installing and configuring it is quite simple:
    - apt-get install ceph
    - vi /etc/ceph/ceph.conf
    - mkcephfs -c ceph.conf --allhosts --mkbtrfs
  - Btrfs is already there

CEPH: FIRST TEST AND FEEDBACK

- Test results:
  - Fully posix compliance (typical home directory usage) OK
  - Performance for small files OK
    - comparable with standard NFS mount point
  - Performance on typical HEP application => WORK IN PROGRESS



CEPH.CONF

```
[global] user = setupuser
; where the mdses and osds keep their secret
encryption keys
keyring = /data/keyring.$name

; monitors
[mon]
;Directory for monitor files
mon data = /data/mon$id
[mon0]
host = node0
mon addr = 192.168.0.100:6789
[mon1]
host = node1
mon addr = 192.168.0.101:6789
[ . . .}


[mds]
[mds0]
 host = node0
[mds1]
host = node1
[mds2]
host = node2


;OSDs
[osd]
; osd data is where the btrfs volume will be
mounted;
; it will be created if absent
osd data = /data/osd$id
; osd journal is the regular file or device to be used
for journaling
osd journal = /dev/sdb2
; The 'btrfs devs' partition will be formatted as
btrfs. btrfs devs = /dev/sdb1
host = node$id
[osd2]
[osd3]
[ . . .}
[osd9]
```

CEPH: CURRENT STATUS

- snapshots — RBD supports read-only named snapshots (and rollback)
- scalable — disk device can be arbitrarily sized (and resized)
- “thin provisioning” — space isn’t used in the cluster until you write to it
- osd: use new btrfs snapshot ioctls (2.6.37), parallel journaling
- mds: clustering, replay fixes
- mon: better commit batches, lower latency updates
- ceph: new gui (ceph -g)

CEPH: CURRENT STATUS

- Since the Ceph kernel client was pulled into Linux kernel 2.6.34, interest in Ceph has greatly increased.
- Ceph is currently the only open source (LGPL licensed) parallel file system that offers a distributed metadata service that is linearly scalable to at least 128 metadata service nodes, supports the POSIX I/O API and semantics, and is able to expand and contract with low overhead without interrupting service.

CEPH: CURRENT STATUS

- CEPH has already plugins to interact Amazon's S3 and Hadoop so it can be deployed in virtual environments such as Amazon's EC2 cloud service, where frequent and significant cluster size changes are the norm.
- Overall Ceph addresses a number of shortcomings of HDFS, i.e., HDFS's limited name-node scalability, its heartbeat overhead, and its highly specialized file access semantics.
- As we write this, Ceph is still experimental and officially not yet ready for production environments. Sage Weil, Yehuda Sadeh, and Gregory Farnum are working full-time on making Ceph production-ready, with new releases coming out every 2 to 4 weeks.