NON C'È CLOUD SENZA STORAGE

CEPH - distributed object storage system

Chi sono?

Nome: ***Dimitri Bellini***

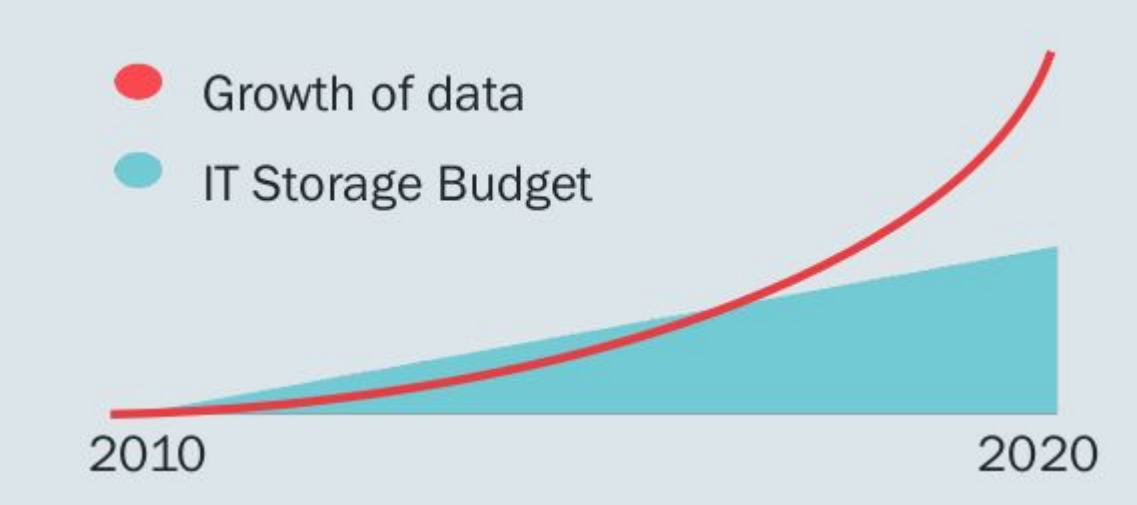
Biografia: *Decennale esperienza su sistemi operativi UX based, Storage Area Network, Array Management e tutto cio' che e' informatica, Official Zabbix Trainer*

Azienda: **Quadrata di Bellini Dimitri**

Profilo Aziendale: *Supporto e consulenza nell'ambito enterprise storage e monitoring*

Sito Web: ***www.quadrata.it***

Il problema a cui rispondere...



Gli storage attuali non sono in grado di scalare facilmente

Aumento della complessita e dei costi

Necessita' di investire su soluzioni proiettate per il futuro

Le soluzioni attuali

Legacy Storage Array

EMC²

IBM

hp

Open Source

DRBD®

Sheepdog

GLUSTER

Evoluzione dello storage

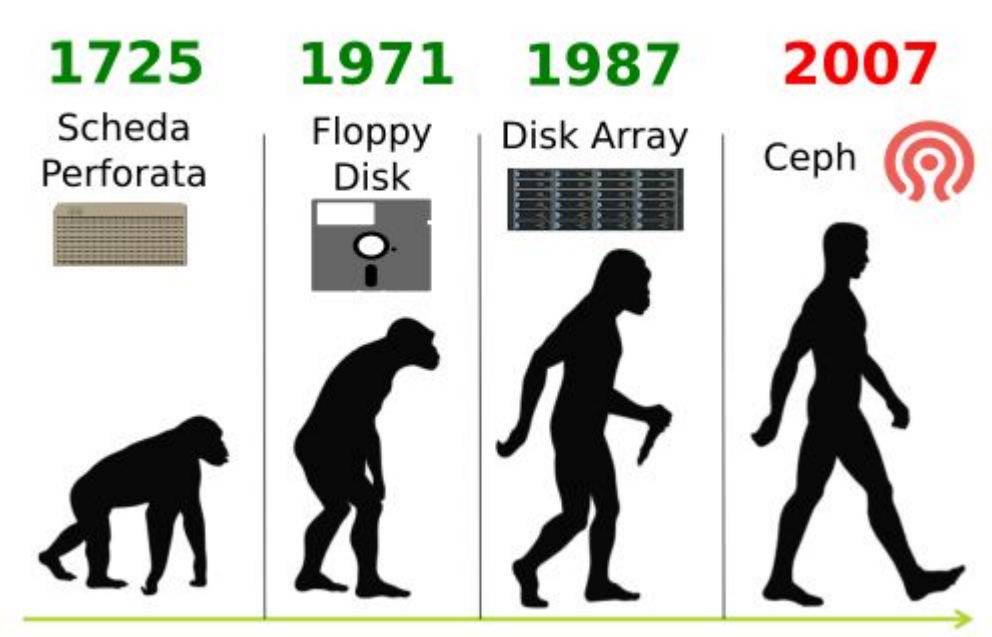
Dalla scheda perforata all’object storage

2007 - nasce **Ceph Opensource Object Storage**

1987 - Nasce il concetto di ARRAY

1971 - Primo Floppy Disk

1725 - nasce la prima forma di archiviazione la “scheda perforata”



La fine dell’epoca RAID?

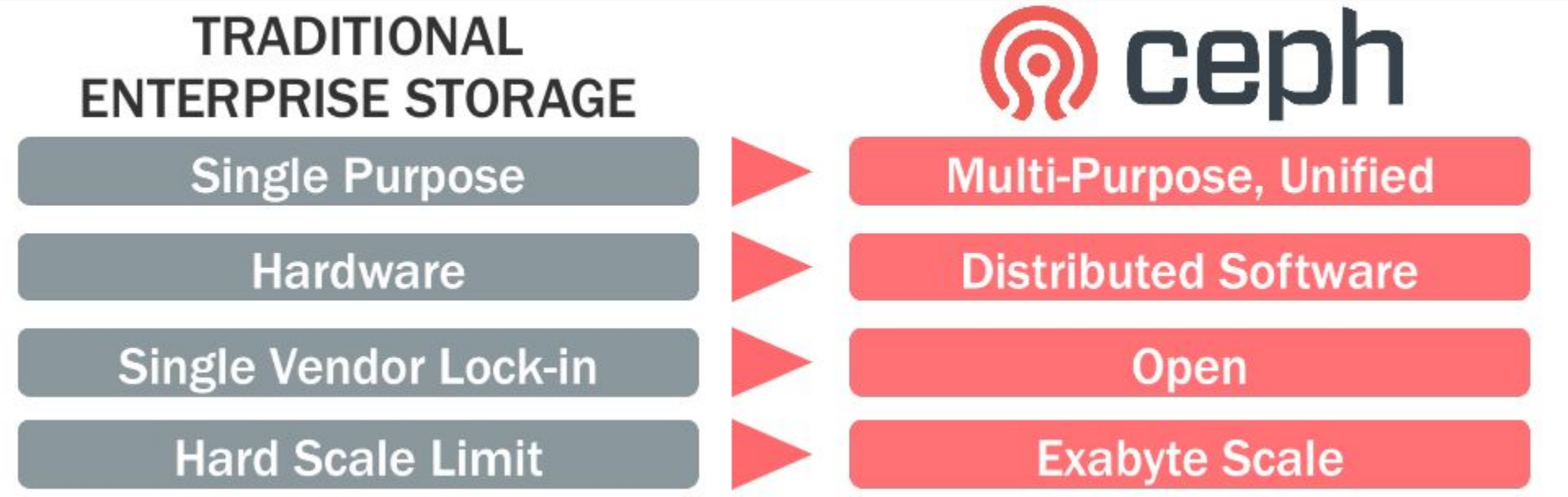
RAID: Redundant Array of Inexpensive Disks

- Enhanced Reliability
  - RAID-1 mirroring
  - RAID-5/6 parity (reduced overhead)
  - Automated recovery
- Enhanced Performance
  - RAID-0 striping
  - SAN interconnects
  - Enterprise SAS drives
  - Proprietary H/W RAID controllers
- Economical Storage Solutions
  - Software RAID implementations
  - iSCSI and JBODs
- Enhanced Capacity
  - Logical volume concatenation

CEPH - Che cos’e’ ?

- **Ceph** è **altamente scalabile**, **open source**, sistema storage di tipo software-defined che puo’ essere installato su **commodity hardware** (comuni server).
- **Ceph** fornisce **object, block** e **file system** storage in unica soluzione self-managing, self-healing senza **nessun single point of failure**.
- **Ceph** puo’ sostituire le soluzioni storage “legacy” e fornisce una soluzione unica per il Cloud.

CEPH vs Soluzioni Tradizionale?

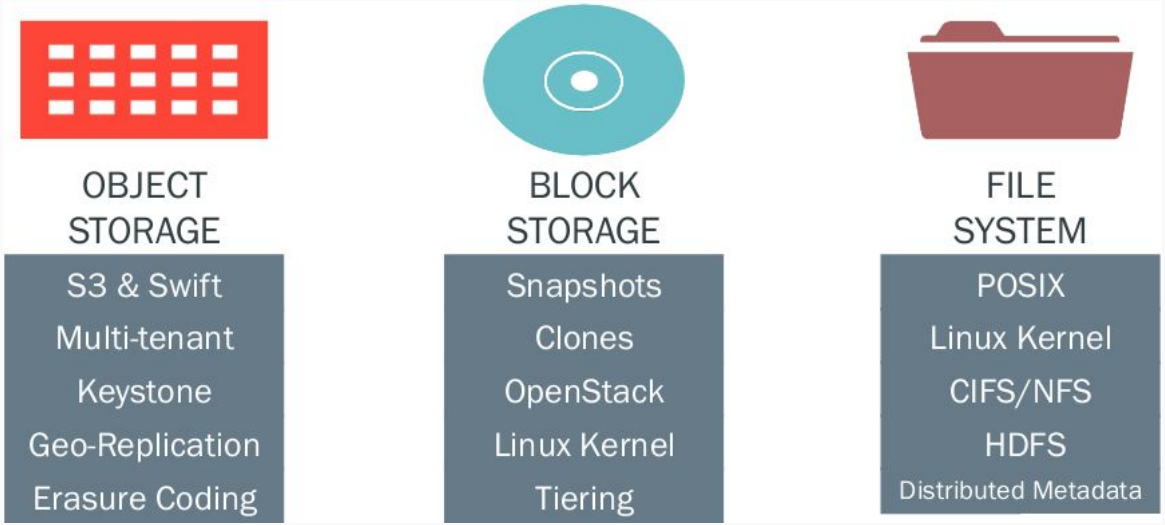


CEPH
Caratteristiche

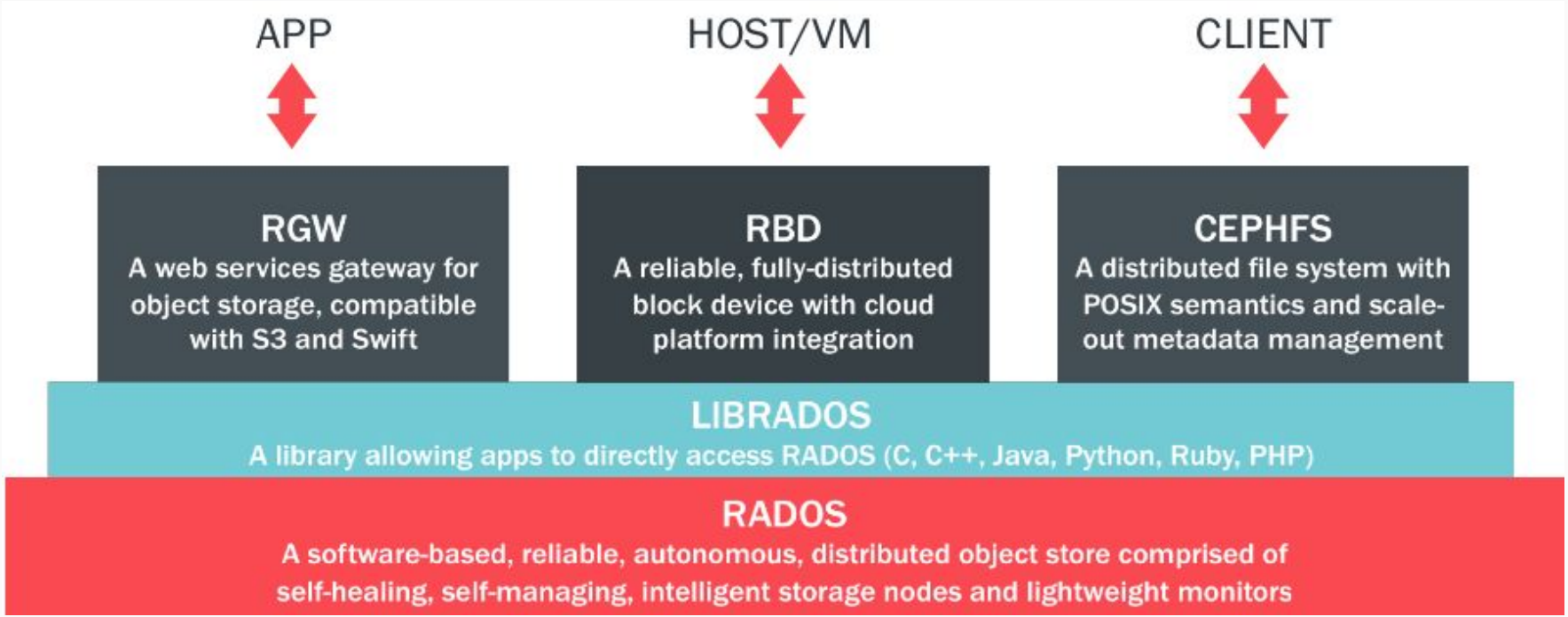
RBD -> Erogazione spazio disco per ambienti CLoud based (KVM, VMWare,etc..)

RGW -> Compatibile S3 standard e SWIFT (Amazon,Openstack), scrittura ad oggetti

CEPHFS -> File System distribuito ideale come sistema NAS



Elementi architetturali



RADOS Daemon (Reliable Automatic Distributed Object Store)

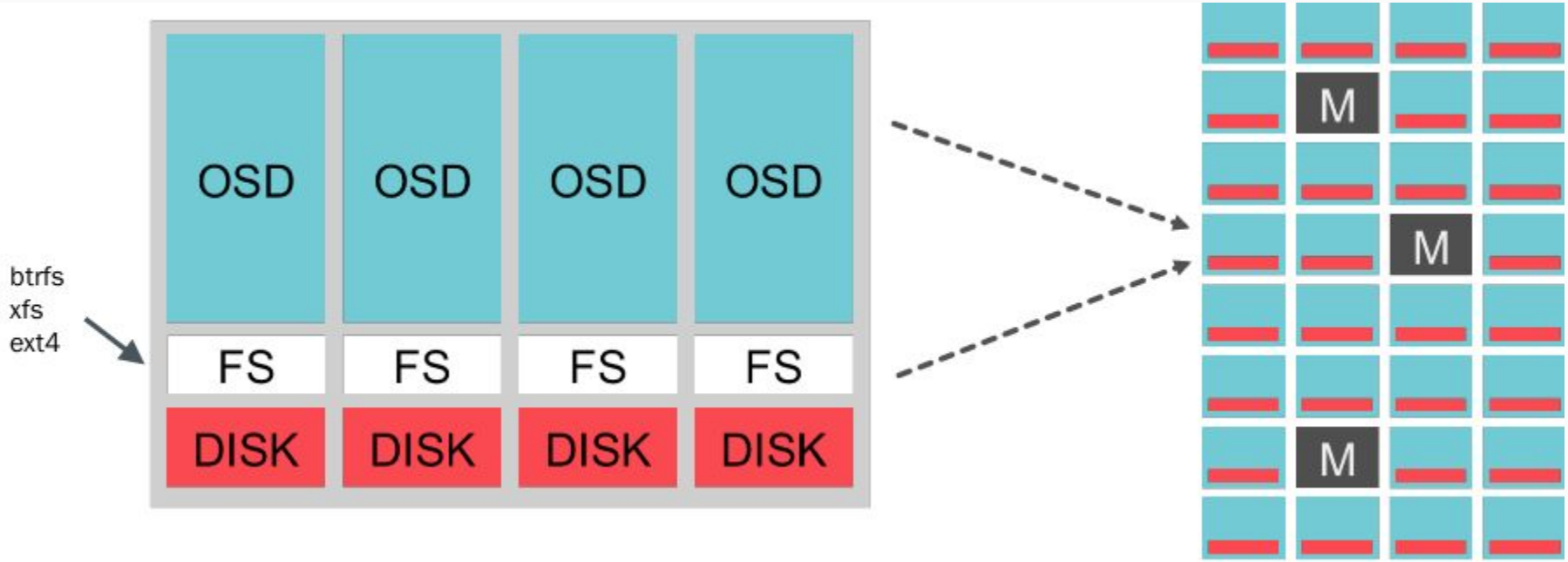
OSDs:

- 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
- Provide consensus for distributed decision-making
- Small, odd number
- These do not serve stored objects to clients

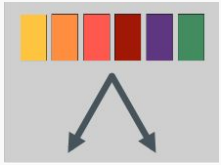
Object Storage Daemon - Dettaglio



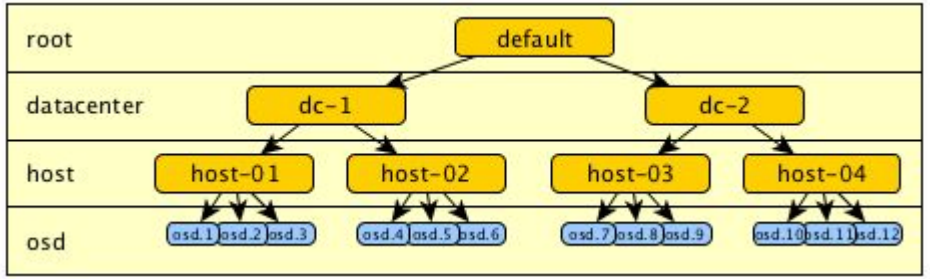
Chi organizza i dati in CEPH?

CRUSH:

- 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

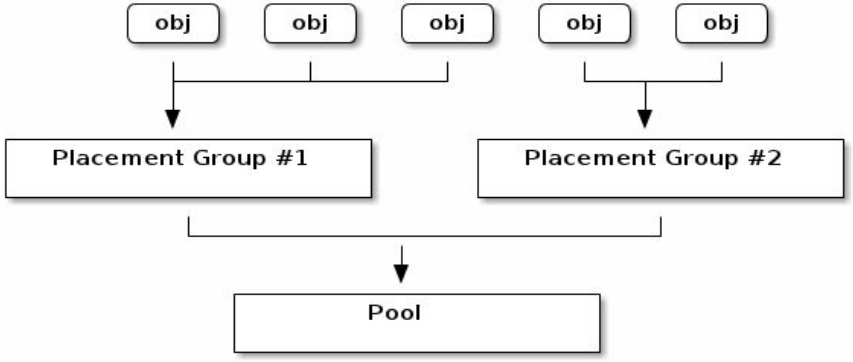


Esempio di CRUSH Map:



Object e Placement Group (PG)

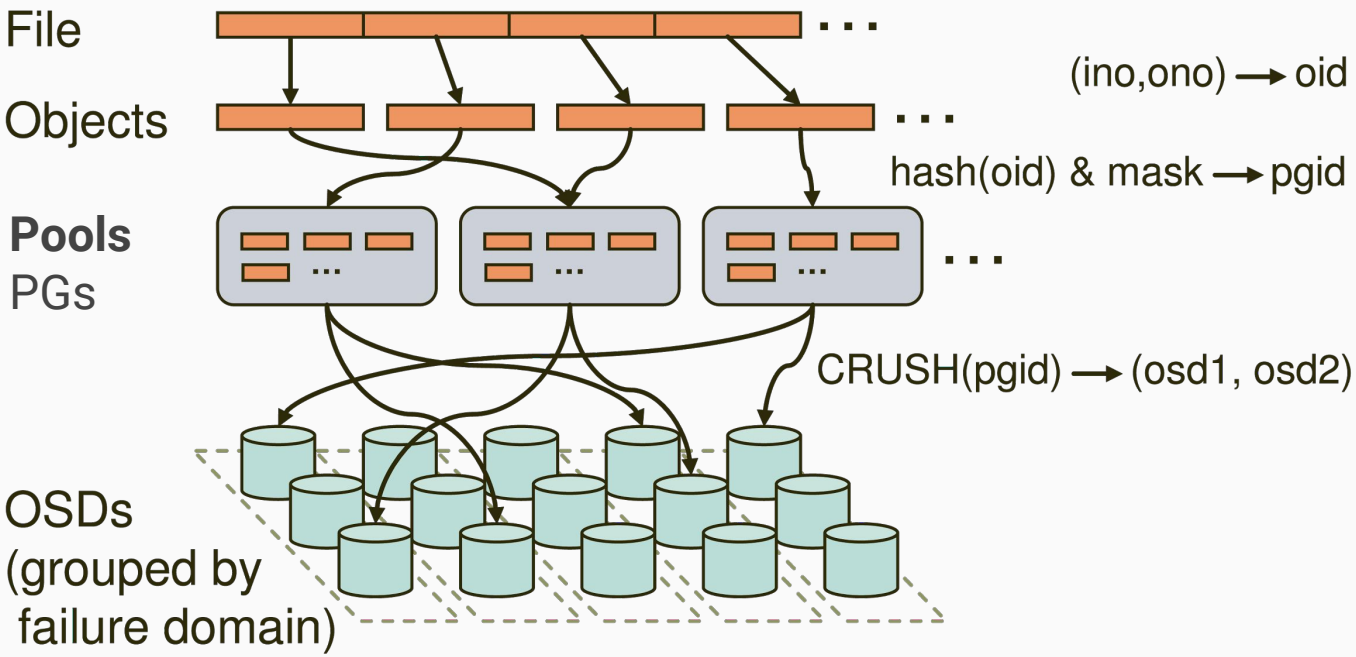
Object : L'Object è la più piccola unità di memorizzazione dati (4MB) in cluster di Ceph, ***tutto viene memorizzato sotto forma di oggetti***. Gli object sono mappati tramite PG e questi oggetti o loro copie sono sempre distribuiti su diversi OSD.



PG (Placement Group): L'algoritmo **CRUSH** associa ogni **Object** ad un **Placement Group** e poi associa ogni **Placement Group** ad uno o più Ceph OSD Daemon. Questo livello di riferimento indiretto consente a Ceph di riequilibrare in modo dinamico quando nuovi Ceph OSD Daemon vengono aggiunti o rimossi.

Con una copia della cluster map e tramite l'algoritmo CRUSH, il client può calcolare esattamente quali OSD sono da utilizzare durante la lettura o la scrittura di un object particolare.

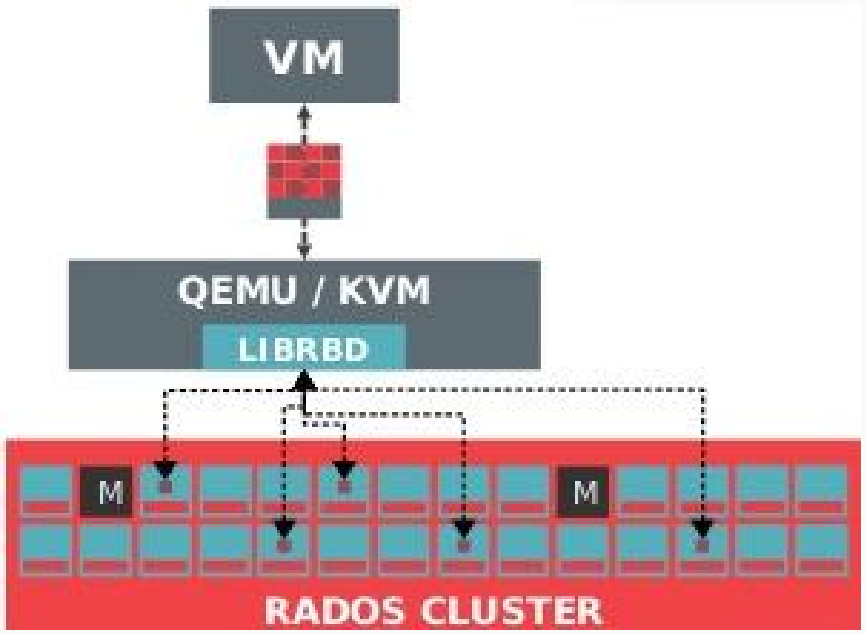
Distribuzione del dato sul cluster CEPH



Ceph & KVM

Tramite le LIBRBD (RADOS BLOCK DEVICE) e' possibile erogare "dischi" virtuali a QEMU/KVM, com questa soluzione il cluster Ceph e' in oltre in grado di garantire alle VM:

- High Availabity (il dato puo' provenire da piu' nodi CEPH)
- Snapshot
- Cloni
- Asynchronous Replication



CEPH Vantaggi: Riorganizzazione PG

Prendiamo ad esempio un disco in errore da 2TB in mirror RAID

- Dobbiamo copiare 2TB dal disco soppravvissuto ad uno nuovo
- Il disco sopravvissuto e quello nuovo risiedono sempre sullo stessa zona

Prendiamo due oggetti RADOS clusterizzati sullo stesso nodo primario

- Le coppie sopravvissute sono riorganizzate (su differenti secondari)
- Le nuove copie saranno riorganizzate (sui diversi successori)
- Vengono copiati 10GB da ciascuno dei 200 sopravvissuti a 200 successori
- Sopravvissuti e successori sono in diverse zone

CEPH Vantaggi: Riorganizzazione PG

Vantaggi

- Il recupero è parallelo e 200x più veloce
- Il servizio può continuare durante il processo di recupero
- L'esposizione ad un probabile 2° guasto è ridotto del 200x
- Gestione della rilocazione in base a “zone” da guasti di livello superiore
- Il recupero è automatico e non sono necessari nuovi dischi
- Non sono richiesti dischi di ricambio in standby

Non piu’ dischi SATA/SAS?!

SAS

- Standard form factor
- 2 SAS ports
- SCSI command set
  - data = read (LBA, count)
  - write (LBA, count, data)
  - LBA :: [0, max]
  - data :: count * 512 bytes
  - CRC on cmd and PI on block

versus

Kinetic Open Storage

- Standard form factor
- 2 Ethernet ports (same connector)
- Kinetic key/value API
  - value = get (key)
  - put (key, value)
  - delete (key)
  - key :: 1 byte to 4 KiB
  - value :: 0 bytes to 1 MiB
  - HMAC on cmd and SHA on value

Ceph e’ la Soluzione Definitiva ???

Lascio a voi provare e verificare se CEPH puo’ soddisfare le vostre esigenze.
Rincordando che **CEPH** e’ **OPENSOURCE**

Grazie!

DOMANDE?

Riferimenti Utili

- www.ceph.com
- www.sebastien-han.fr
- karan-mj.blogspot.it