Lesson 3 - Cloud infrastructure – storage and data repositories

Milan Brož
Software engineer
Storage engineering

Agenda

- Storage data center elements
- Layers
- Data protection
- Virtualization
- Distributed storage
- Security
- Performance
- Management and API

Storage – Data center elements

Storage

- Capacity
- Availability, Reliability
- Data integrity, Redundancy
- Performance
- Scalability
- Security

Manageability

=> Cost

SDS – Data center elements

Software Defined Storage (SDS)

- "Comodity hardware with abstracted storage logic"
- Policy-based management of storage
- Virtualization
- Resource management
- Similar concept as Software Defined Network (SDN)

  Note: distributed storage is mostly about networking!
- Thin provisioning, deduplication, replication, snaphsots,

  ...

  SDS definition differs among vendors!

HW – Storage layers

Hardware and low-level storage protocols

- **Physical storage**
  - Rotational drives / hard disk drives (HDD)
  - Flash / SSD drives
  - Persistent Memory (byte-addressable!)
  - Tapes, magneto-optical drives, ...

- **Block-oriented storage access protocols**
  - "Small Computer System Interface" (SCSI), Serial Attached SCSI (SAS)
  - Serial ATA (SATA)
  - Fibre channel (FC) (not only fiber-optic)
  - InfiniBand (IB)

NAS – Storage layers

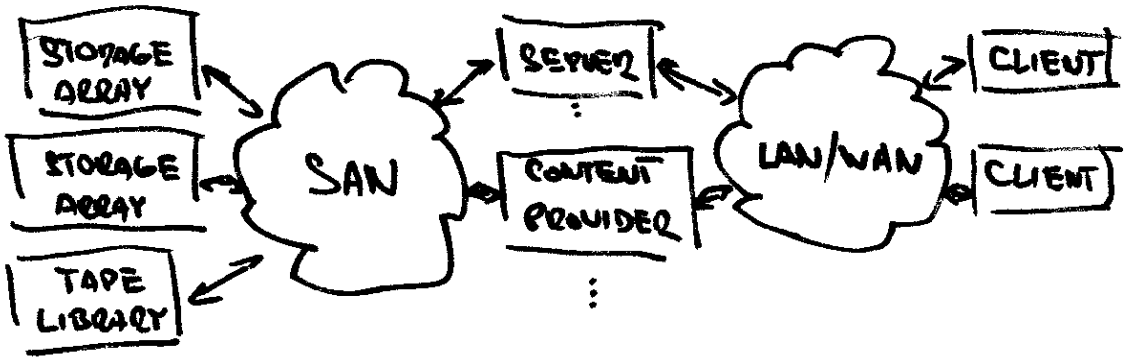
Storage connectivity through network

- **Direct-Attached Storage (DAS)**
  - local, host-attached
- **Network-Attached Storage (NAS)**
  - remote storage device
  - communication protocol
    - usually over IP-based network
    - high-level: NFS, CIFS, HTTP, ...
    - low-level: iSCSI (SCSI over IP), FC (point-to-point), Network Block Device (NBD)

SAN – Storage layers

Storage connectivity through network

- **Storage Area Network (SAN)**
  - private network
  - switched fabric
  - communication protocol
    - Fibre Channel
    - InfiniBand
    - FC over Ethernet (FCoE)



Data protection

- **Data integrity protection**
  - random error detection (parity) / correction
- **Erasure codes** – Forward Error Correction (FEC)
  - Redundancy
  - RAID (Redundant Array of Independent Disks)
  - Erasure coding in distributed storage
- **Backup and disaster recovery**
  - **"RAID is not a backup!"**
    - File corruption, bugs (disk, controller, OS, application, ...)
    - Admin error, malware
    - Catastrophic failure (datacentre fire)
  - Offline and off-site backup replica

RAID – Data protection

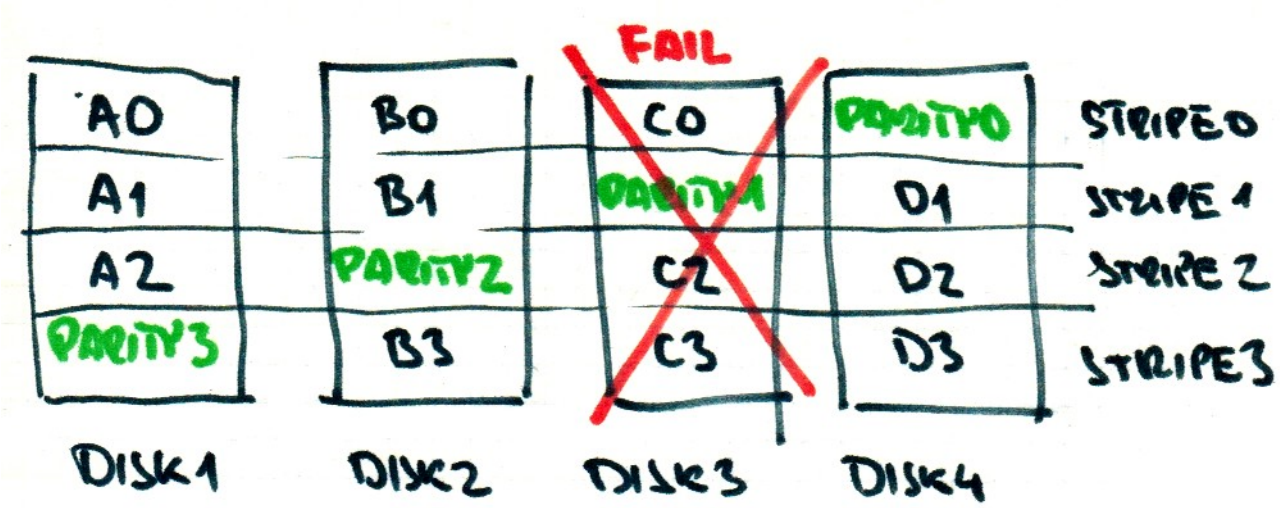
Common non-RAID and RAID disk configurations

- **JBOD** – "Just a Bunch of Disks" (collection of disks, no redundancy)
- **RAID-0** – striping (for performance, no redundancy, no parity)
- **RAID-1** – mirroring (no parity)
- **RAID-5** – block-level striping + distributed parity (XOR)
- **RAID-6** – block-level striping + double distributed parity
- **RAID-10** – nested RAID example (1+0: striping over mirrored drives)
- **RAIDZ** (in ZFS) – similar to RAID-5, dynamic stripes, self-healing
- **MAID** (Massive Array of Idle Disks) – "Write once, read occasionally"

  ...
- **Degraded mode**
  - RAID-5 (RAID-6 soon): large drives reconstruction time, fail during rebuild
- **Hardware RAID** vs **software RAID** vs **"fake RAID"** (processing in fw/driver)

RAID – Data protection

RAID-5 schema + example of disk fail (**"erasure"**)



Erasure coding – Data protection

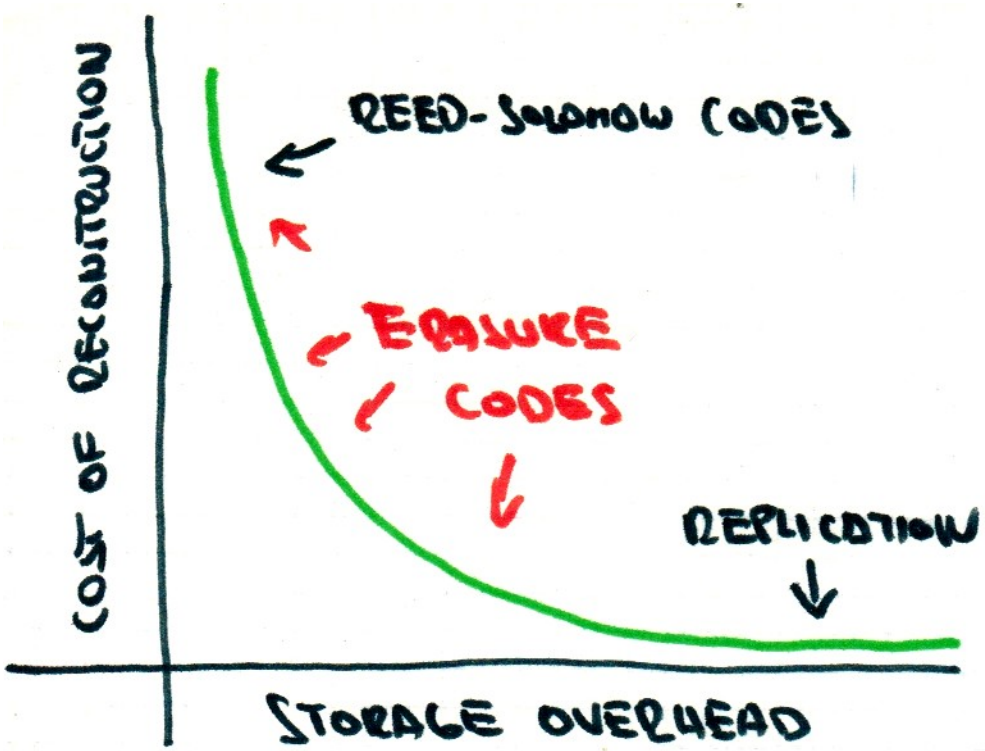
- **Data protection is trade-off**
  - Storage overhead
  - Reconstruction cost
  - Reliability

  - Still active research ...
    - From simple XOR (RAID) to Galois Field arithmetic – GF(2^x)
    - Reed-Solomon codes, Pyramid codes
    - Bit-Matrix codes

      ...

Erasure coding – Data protection

Erasure codes trade-off and efficient solution

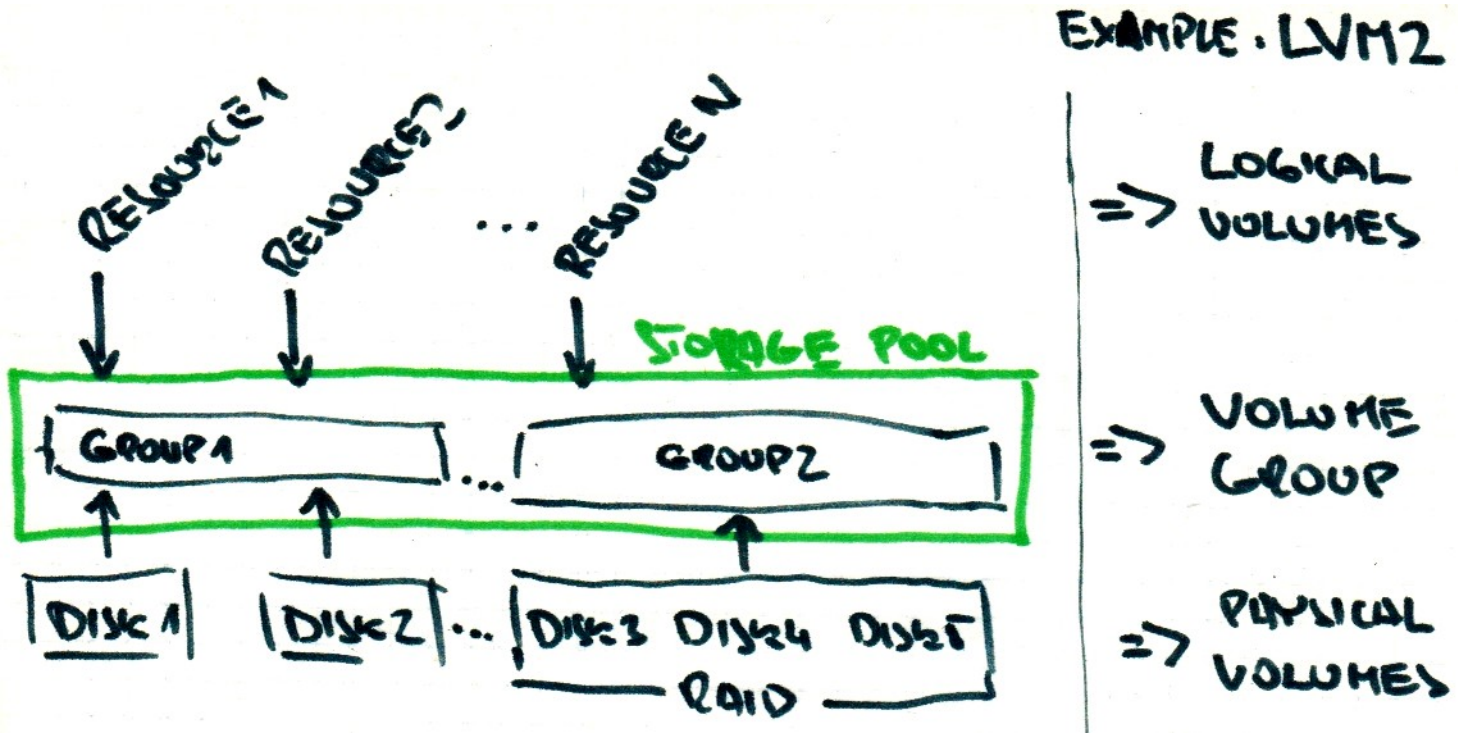


Storage Pool – Virtualization

- **Storage pool**
  - set of disks, blocks, ... => allocatable area for data
- **Pre-allocated storage**
  - partition table, logical volume in Logical Volume Manager (LVM)
- **On-demand allocated storage**
  - **Thin provisioning** (only blocks in use are allocated)
  - Flexible allocation
  - Used in snapshots
  - Possible over-allocation (sharing "unallocated" space)

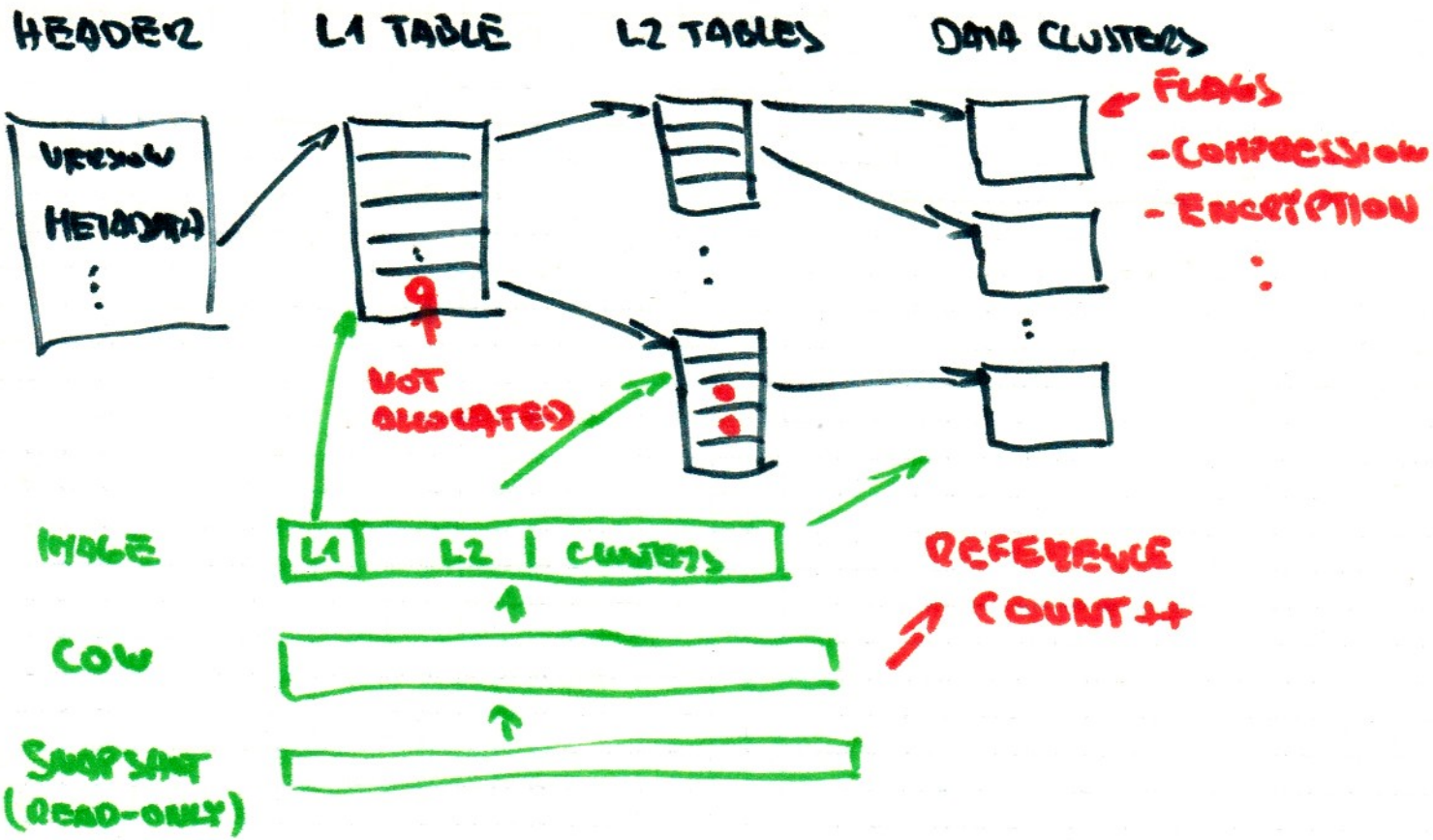
Volume manager – Virtualization

Storage Pool – example for Linux Logical Volume Manager (LVM2)



QCow2 format – Virtualization

QCow2 image format – allocation principles (pre-allocated vs on-demand)



Distributed storage

Distributed object store or network file-system

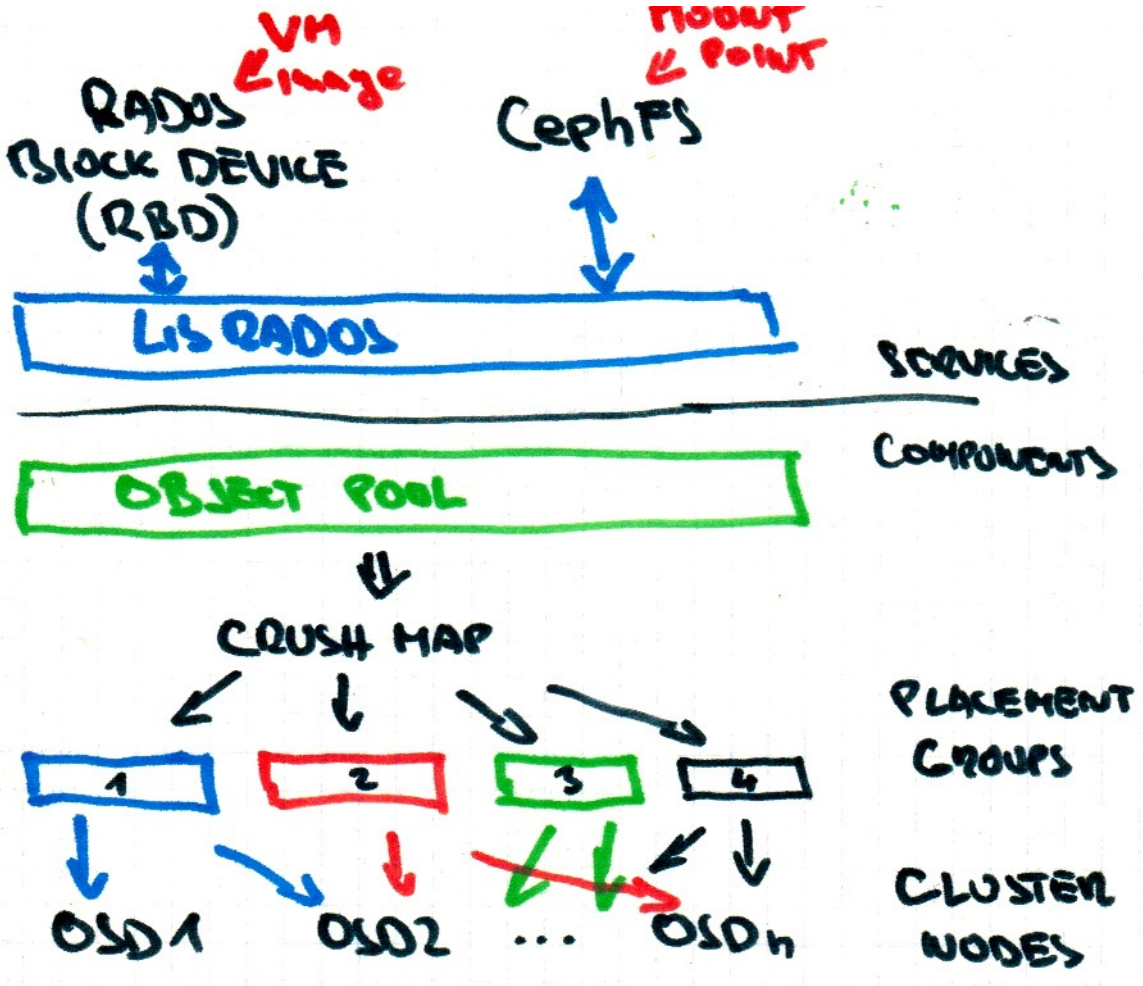
- **Transparency**
  - Access (same as local), Location (any node), Failure (self-healing)
- Backend for Cloud services

- Examples
  - Ceph, GlusterFS (Red Hat)
  - HDFS – Hadoop File-System (Apache)
  - Windows Distributed File-System (Microsoft)
  - GFS (Google)
  - Isilon (EMC^2)

    ...

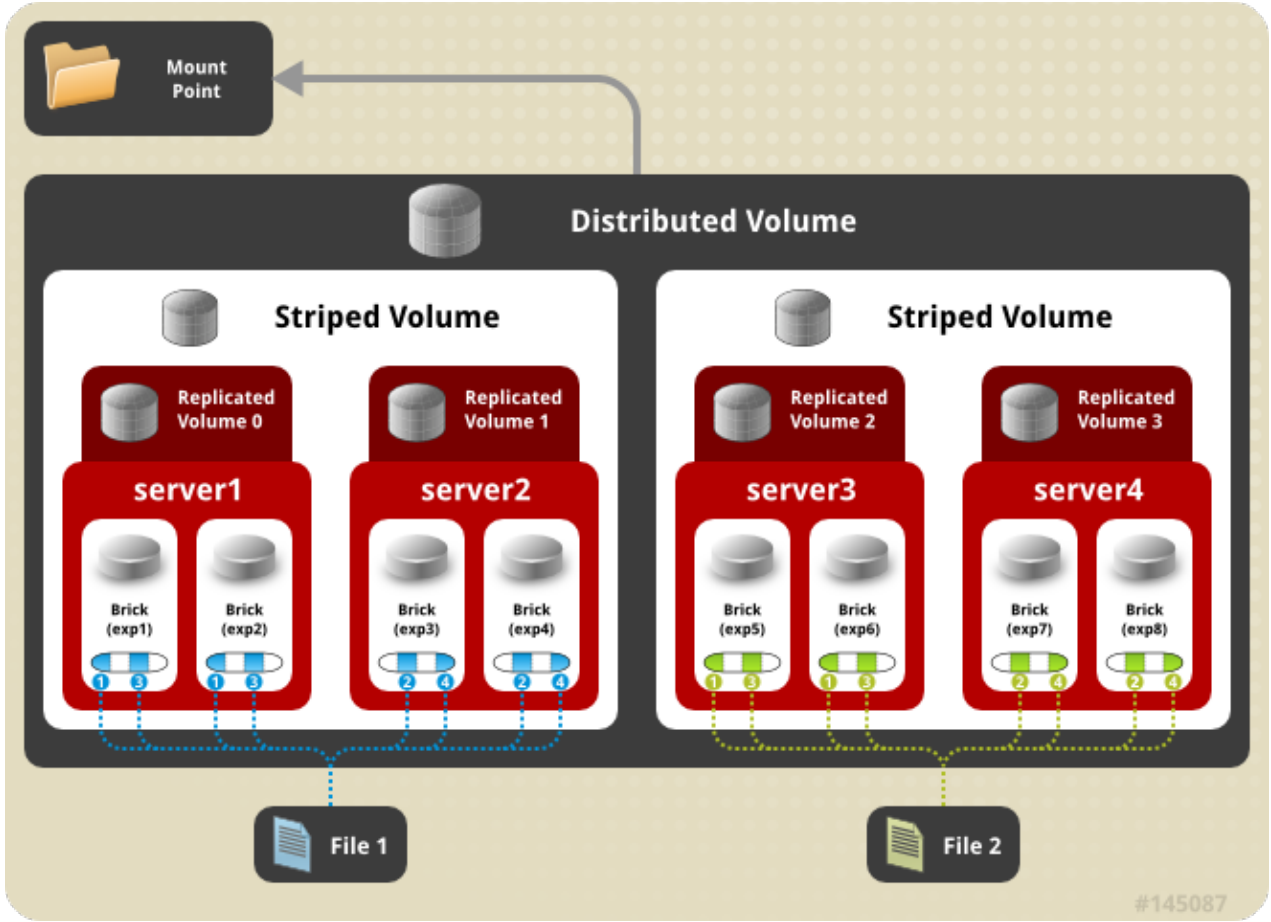
CEPH – Distributed storage

- **CEPH principles**
  - block device (RBD)
  - CephFS
  - object store
  - libRADOS
  - CRUSH
  - Placement group
  - OSD



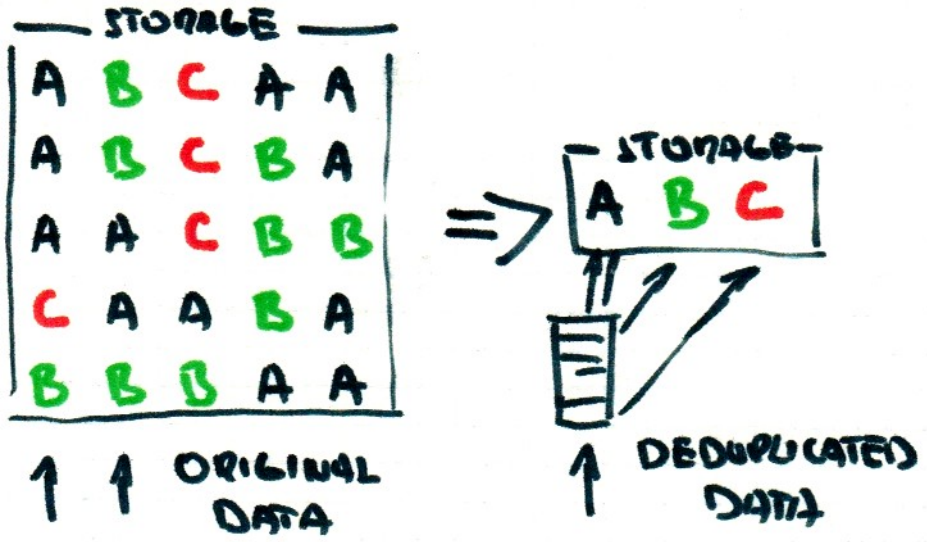
GlusterFS – Distributed storage

Example of access of **GlusterFS** resources



Deduplication / Compression

- **Deduplication**
  - avoid to store repeated data
  - file or block level
  - space-efficient, stateless mode
  - deduplication performance
  - data corruption amplification

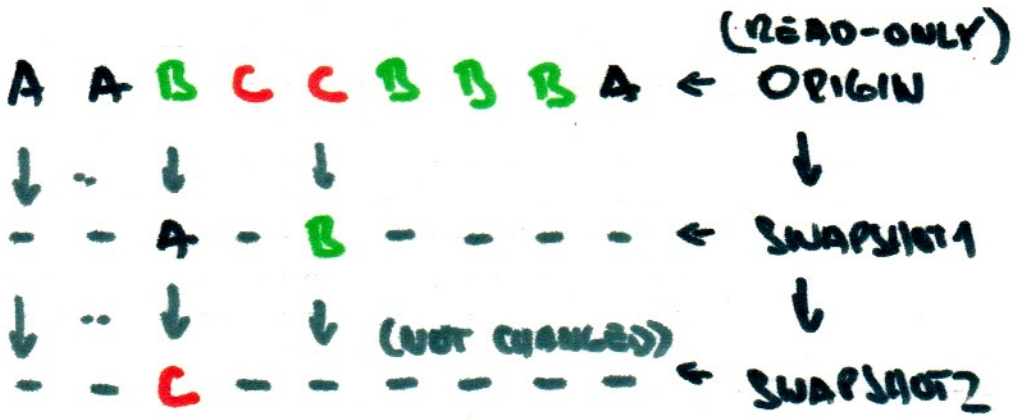


- **Compression**
  - more generic algorithms
  - special case: zeroed blocks

Snapshots

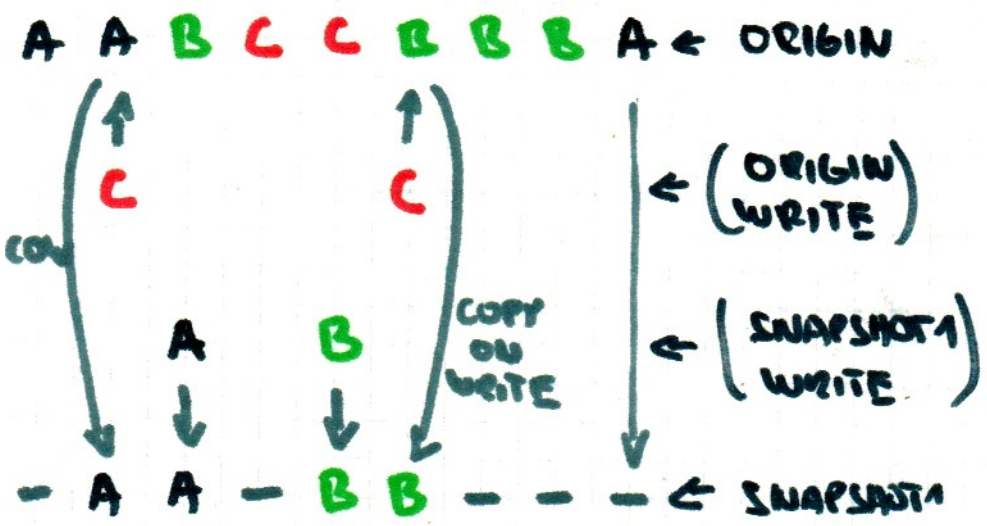
- **Snapshot of storage** in specific time
  - implicit deduplication
  - Allows quick revert to older state (recovery)

- Principle of writable snapshots with read-only origin



Snapshots

- Principle of snapshot with writable origin



- **Copy on Write (COW)** principle
  - delayed copy to snapshot (before origin write)
  - write to origin => need to copy the changed block first

Templates

Template

- Application of deduplication + snapshots (+ thin provisioning)
- **Virtual machine template**
  - base operating system
  - common configuration (networking, firewall, ...)
  - common applications (webservers, user packages, ...)
- One base image, only changes are stored

- Application containers + template
  - used in Docker

Security

- **Security policies**
- **Confidentiality**
  - Storage encryption and data connection encryption
  - Key management
- **Authentication**
- Integrity (in cryptography sense – authenticated encryption)
- **Access control, permissions**
- **Secure data disposal / destruction**
- **Audit**

...

Adding security later in game usually does not work :-)

... common mistake in storage engineering

Performance

- **Tiered storage**
  - Several layers of storage in one chain
  - Different performace, availability, recovery requirements
  - Cache

- **Multi-path** (also High Availability)
  - storage multipath
  - network connection (link bonding)

- Virtualization of drivers
  - virtio, pass-through device

High Availability

- **High availability (HA)** – assuring access to resources (data)
  - Service-level agreement (SLA)
  - common 9s levels

UPTIME (%)	DOWNTIME (%)	DOWNTIME PER YEAR	DOWNTIME PER WEEK
98	2	7.3 days	3 hr 22 minutes
99	1	3.65 days	1 hr 41 minutes
99.8	0.2	17 hr 31 minutes	20 minutes 10 sec
99.9	0.1	8 hr 45 minutes	10 minutes 5 sec
99.99	0.01	52.5 minutes	1 minute
99.999	0.001	5.25 minutes	6 sec
99.9999	0.0001	31.5 sec	0.6 sec

- **HA clusters**
  - several types of clusters
  - cluster resource access
    - on-demand
    - active/passive
    - active/active
  - resource locking / pseudolocking (shared storage access / SANLock)
  - load-balancing

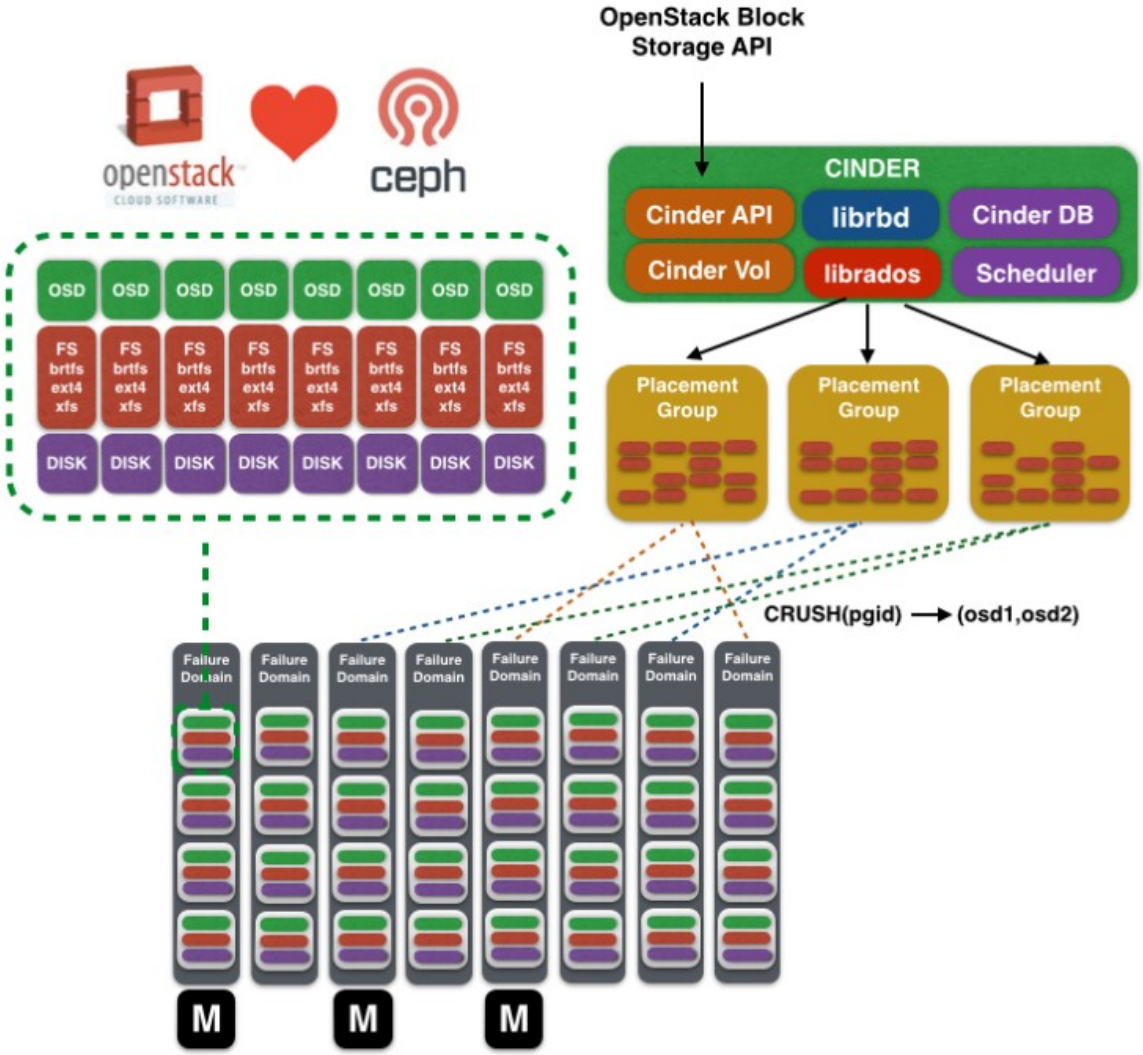
Cloud Storage API

- Open / proprietary API (Application Program Interface)
  - possible vendor lock-in

- Multitenancy
- Cloud – network access
  - RESTful – object-oriented storage
  - JSON / XML

- Example of integration OpenStack & CEPH

OpenStack & CEPH – Storage API



Questions?