redhat

PA200 – Cloud infrastructure
Storage and Data repositories

Milan Brož
mbroz@redhat.com

redhat

Agenda

Storage cost
OpenStack storage types
Software-defined storage concepts
Data persistence and redundancy
Virtualization
Distributed storage
Security
Q & A

Also see: Information Storage and Management, 2nd Edition
EMC Education Services, ISBN: 978-1-118-09483-9

Storage Cost

Storage

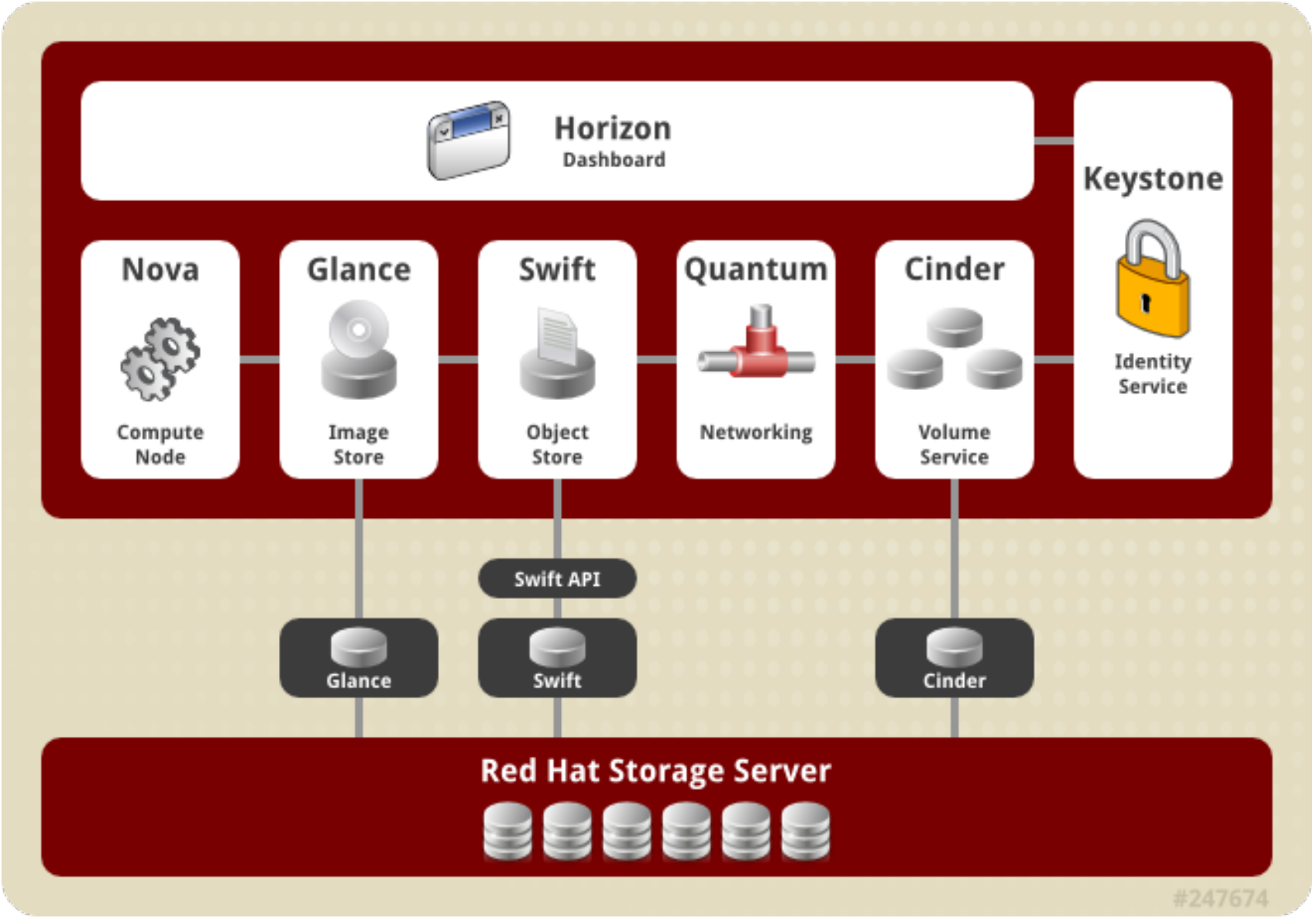
- Capacity
- Availability, Reliability
- Data integrity, Redundancy
- Performance
- Scalability
- Security

=> Cost

Manageability

Storage in OpenStack as an example

Persistent Storage – Example



Data persistence

Ephemeral storage

- Disappears when VM is terminated
- Temporary data ~ computing clusters
- Visible locally (to node)

Persistent storage

- Data always available (no dependency on instance)
- Can be shared among resources / instances

Persistent Storage Types

Object store

- Binary objects of various length
- REST API

Block (volume) storage

- Block (sector-level) devices
- Can be backed by a file image

Shared file-system storage

- Mounted to a directory

Persistent Storage – OpenStack

Object store = SWIFT

Stateless swift-proxy

Block (volume) storage = CINDER

Backend Cinder drivers (LVM, GPFS, EMC, …)

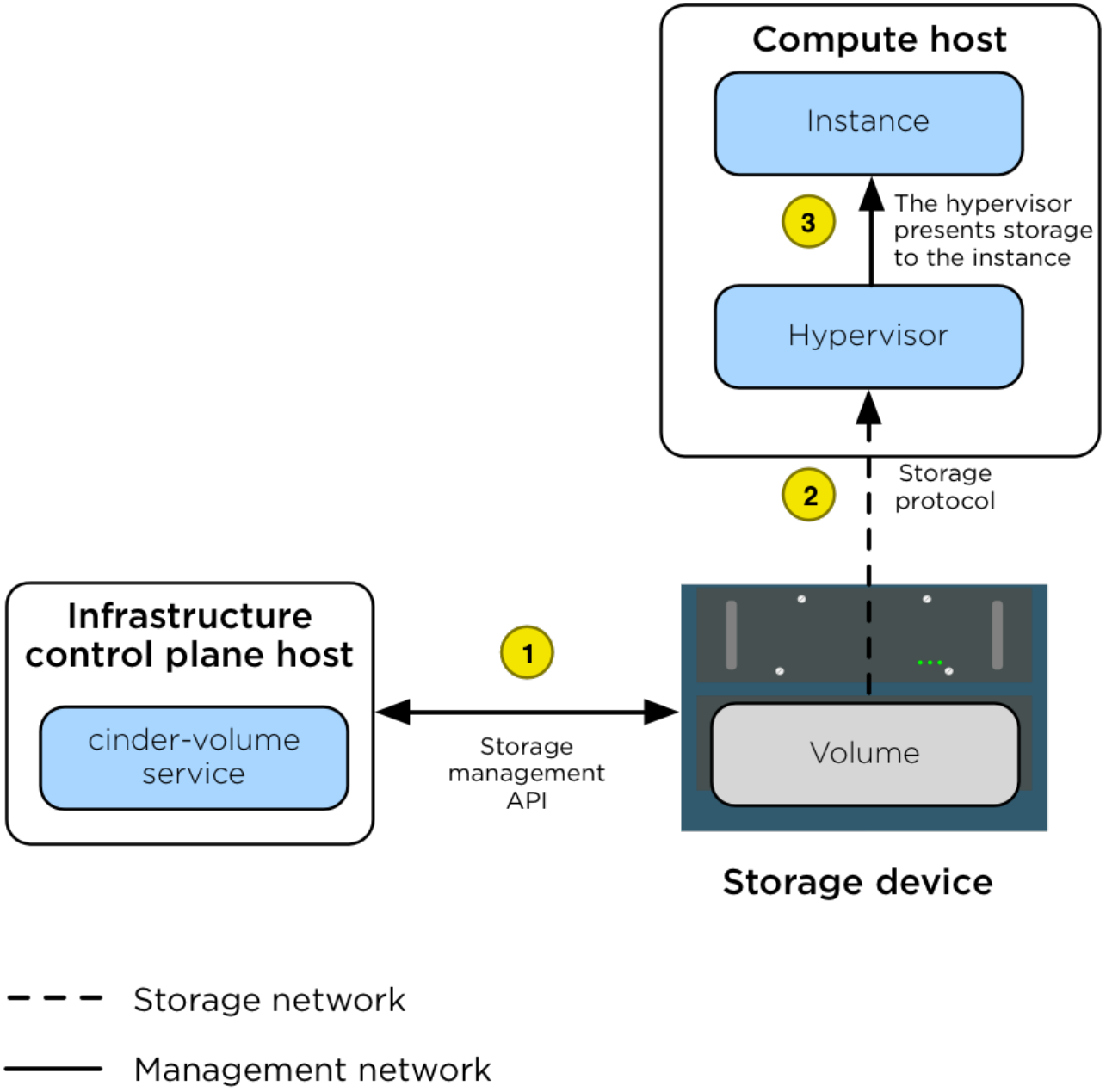
Shared File-system = MANILA

Backend Manila drivers (Ceph, GlusterFS, NFS, …)

Image service = GLANCE

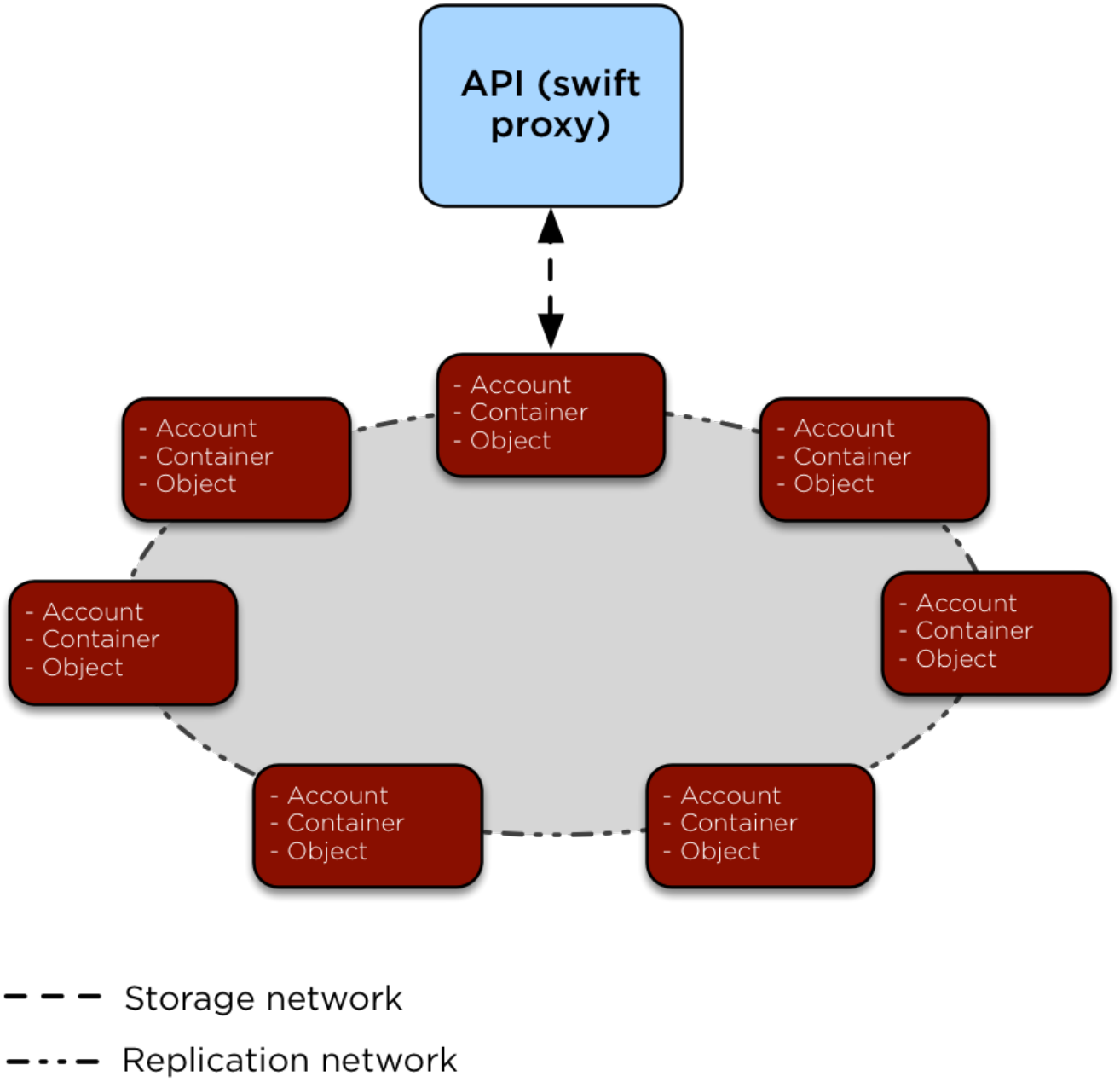
deduplication, clones, ...

Cinder storage overview



https://docs.openstack.org/project-deploy-guide/openstack-ansible/newton/overview-storage-arch.html

Swift storage overview



https://docs.openstack.org/project-deploy-guide/openstack-ansible/newton/overview-storage-arch.html

Generic Storage Concepts

Software Defined Storage (SDS)

- "Commodity 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, snapshots, ...

  SDS definition differs among vendors!

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

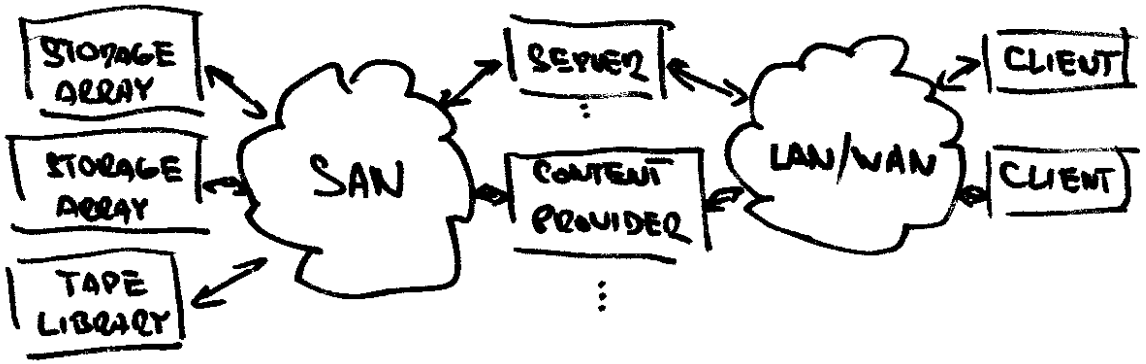
Storage connectivity through network

- **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)
  - Multipath



High Availability

High availability (HA)

- Assuring access to resources
- 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

Resources access

- On-demand
- Active/Passive
- Active/Active

~ Active/Passive – Mid-Range

~ Active/Active – High-End

Generic Storage Concepts
Data Protection and Redundancy

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

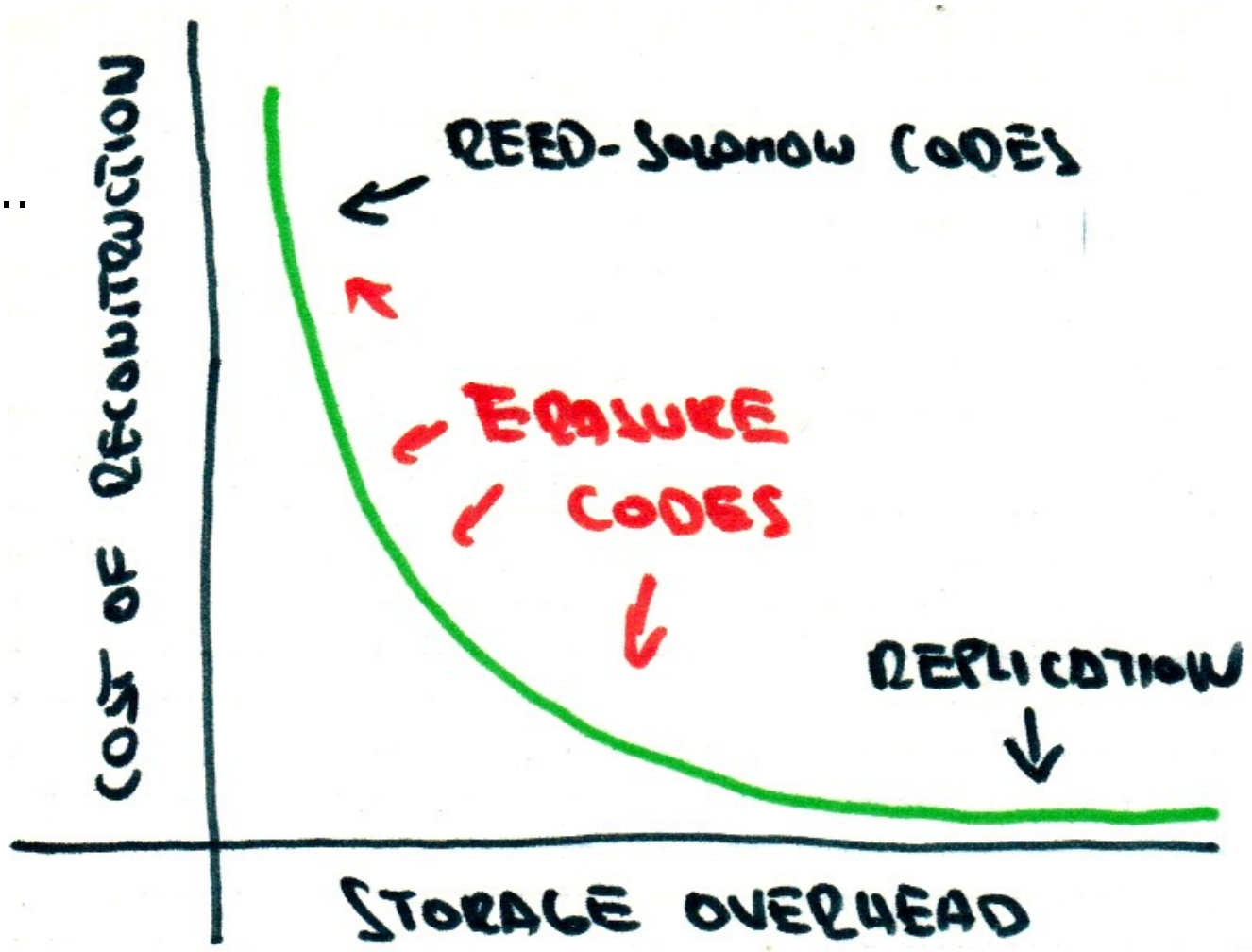
RAID – Data protection

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 reconstruction time, fail during rebuild
- **Hardware RAID** vs **software RAID** vs **"fake RAID"** (in fw/driver)

Erasure coding – Data protection

- **Data protection is trade-off**
  - Storage overhead
  - Reconstruction cost
  - Reliability
  - Still active research ...



Generic Storage Concepts Virtualization

Storage pool

Set of disks, blocks, ... allocatable area for data

Pre-allocated

- Partition table, logical volume in Logical Volume Manager

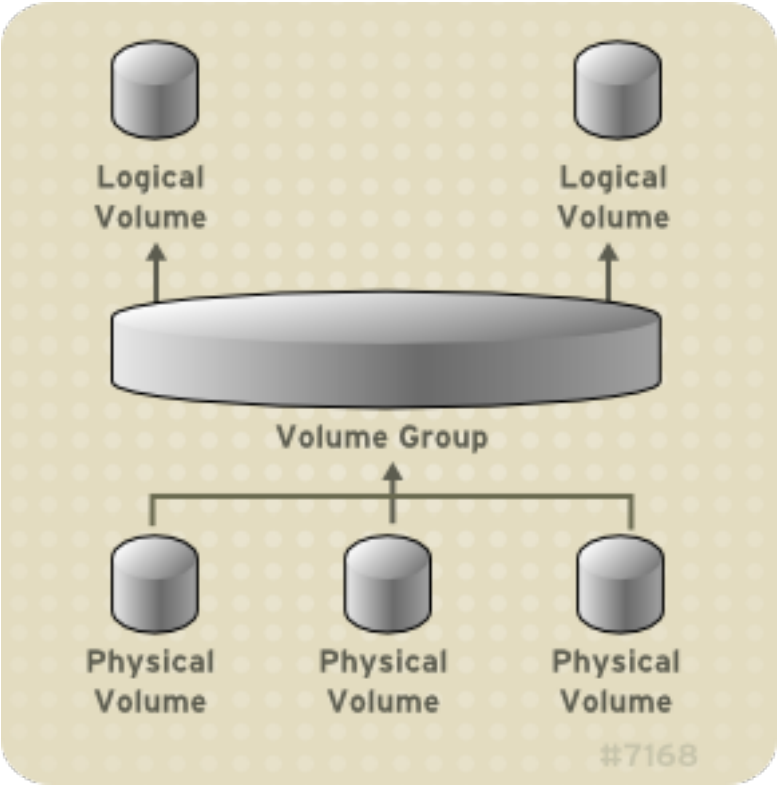
On-demand allocated

- **Thin provisioning** (only blocks in use are allocated)
- Flexible allocation
- Used in snapshots
- Possible over-allocation (sharing "unallocated" space)

Storage Pool Example

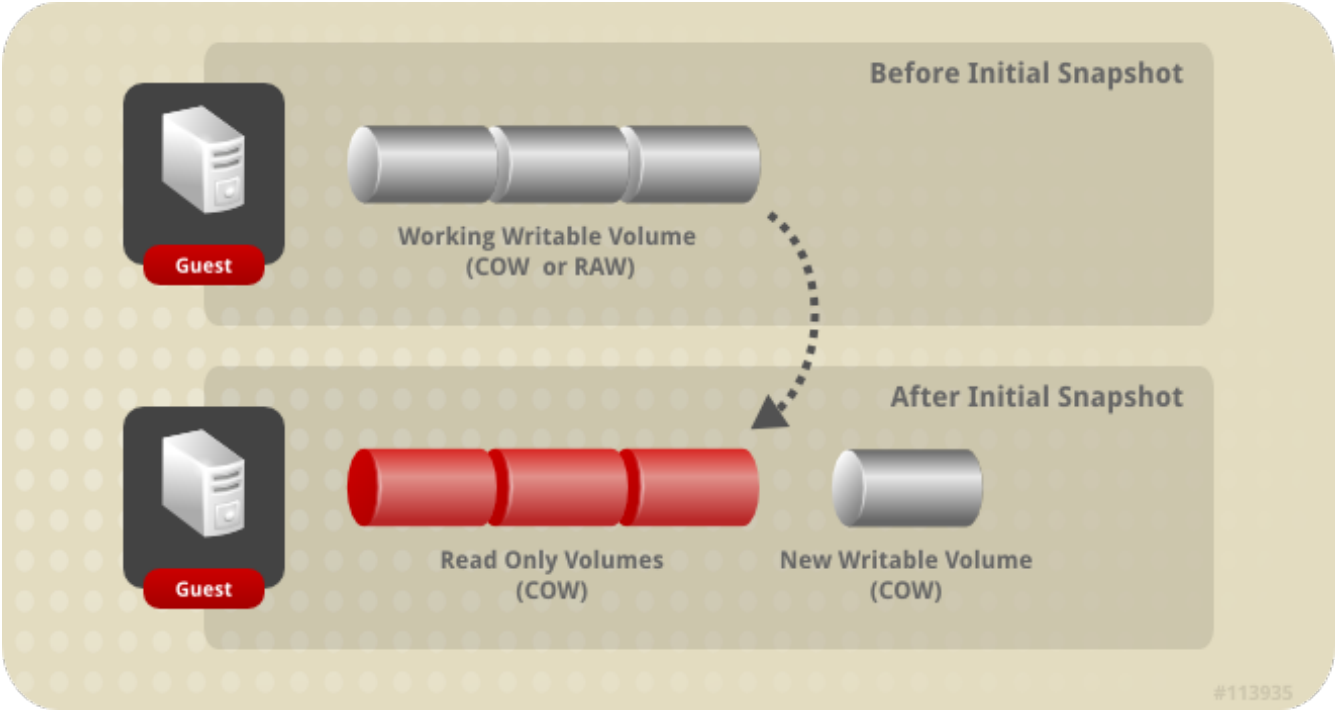
Volume Group

Logical Volume Manager (LVM)



Snapshots

- **Snapshot of storage** in specific time
  - Allows quick revert to older state (recovery)
- **Copy on (first) Write (COW/COFW)** principle
  - Delayed copy to snapshot (before origin write)
  - Write to origin => need to copy the changed block first



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

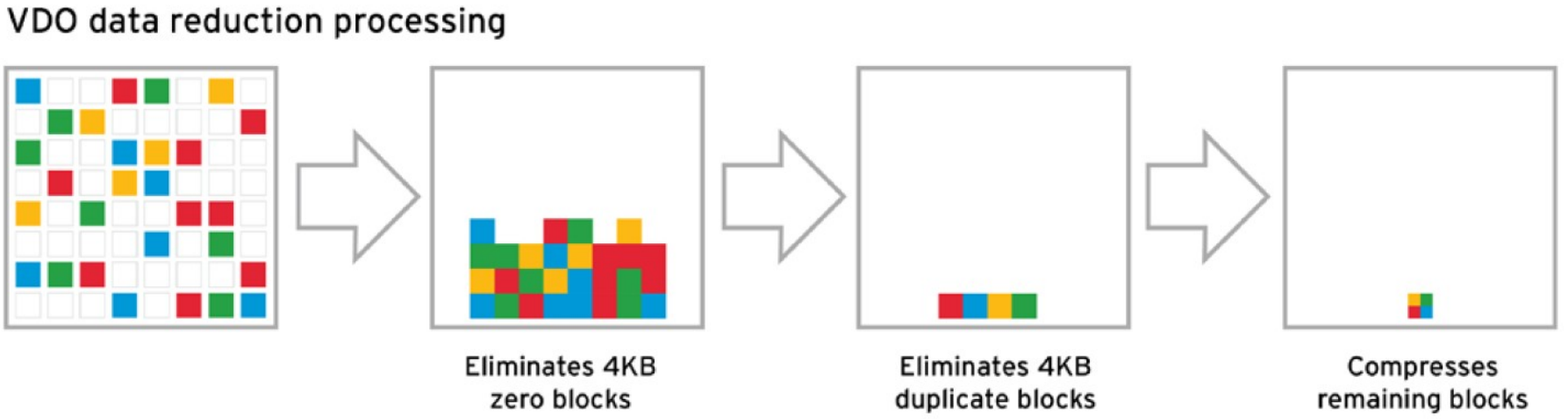
Deduplication / Compression

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



Performance

Tiered storage

- Several layers of storage in one chain
- Different performace, availability, recovery requirements
- Cache (REST API)

Virtualization of drivers

- virtio, pass-through device

Generic Storage Concepts
Distributed Storage

Distributed storage

Clustered

- Cooperating nodes

Distributed

- Storage + network

Distributed storage transparency

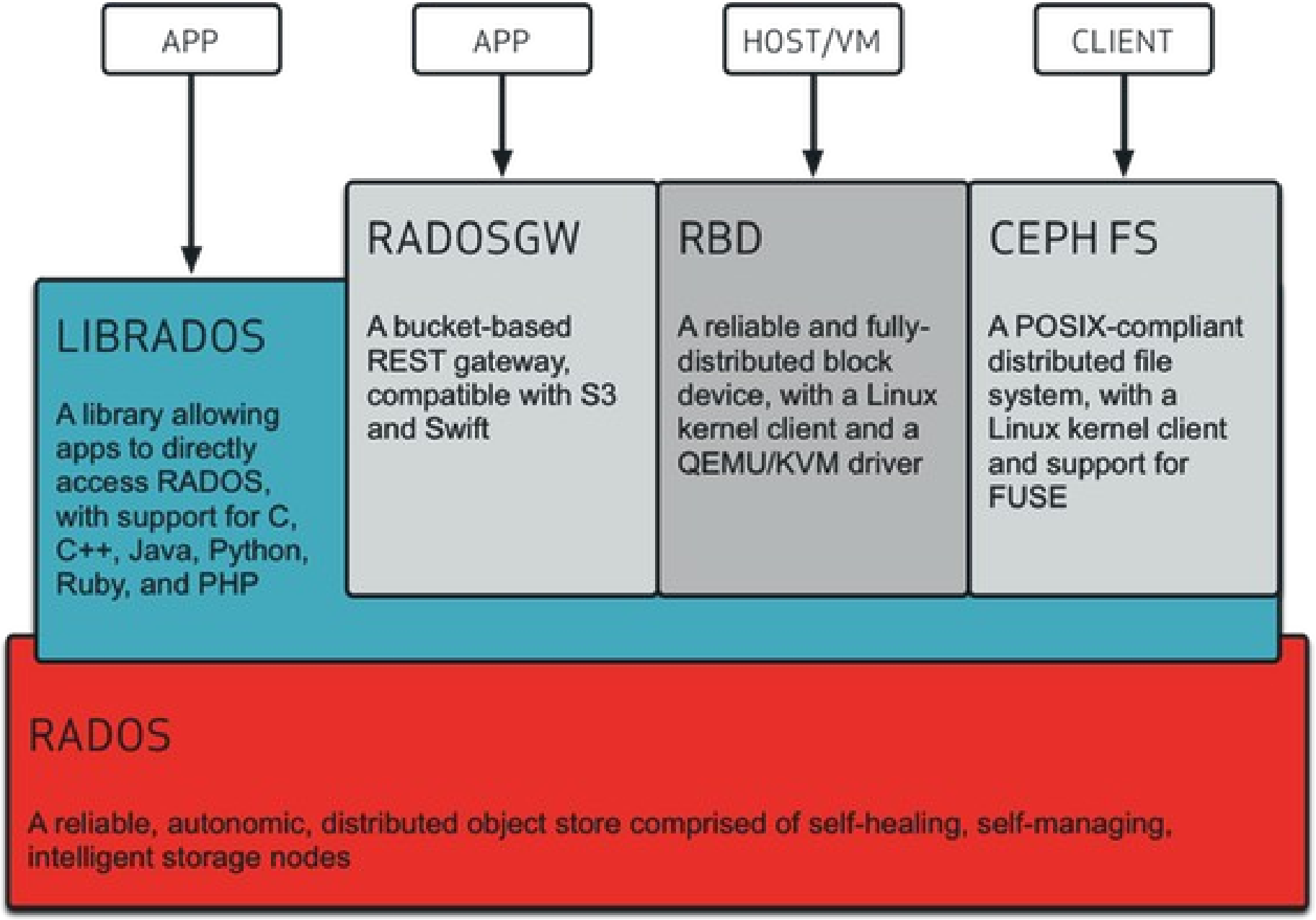
- Access (same as local)
- Location (any node)
- Failure (self-healing)

Distributed storage examples

- Ceph, GlusterFS (Red Hat)
- General Parallel File System – GPFS (IBM)
- Hadoop File-System HDFS (Apache)
- Windows Distributed File-System (Microsoft)
- GoogleFS / GFS (Google)
- Isilon (EMC^2)

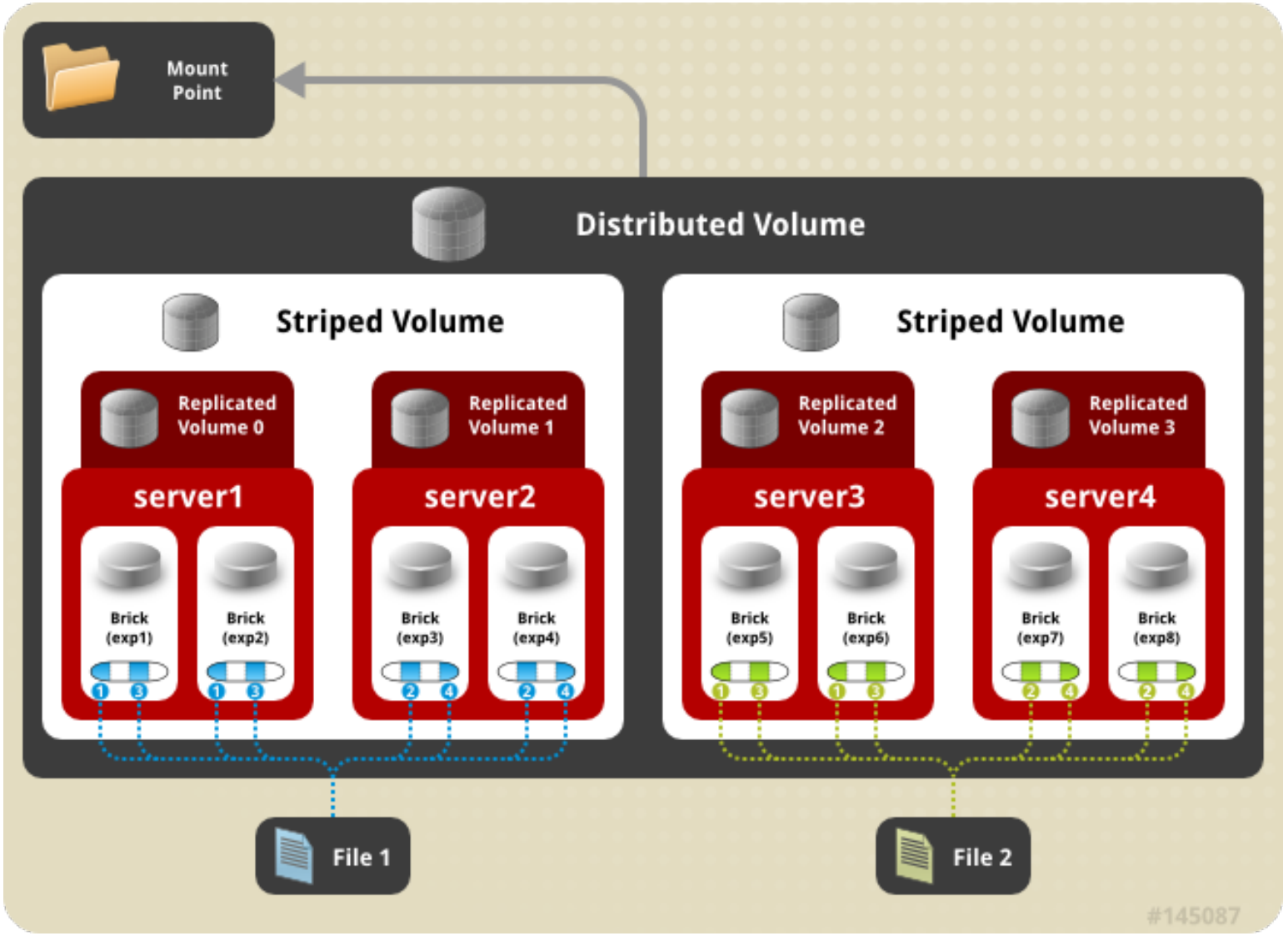
...

CEPH – Distributed storage



GlusterFS – Distributed storage

Example of access of **GlusterFS** resources



Generic Storage Concepts Security

Security

- **Security policies**
- **Confidentiality**
  - Storage encryption (at-rest)
  - Data connection encryption (in-transit)
  - Key management
- **Authentication**
- **Integrity** (authenticated encryption)
- **Access control, permissions**
- **Secure data disposal / destruction**
- **Audit**

...

Cloud Storage Encryption

Encryption on client side

- "End-to-End" encryption
- Lost Efficiency for deduplication/compression

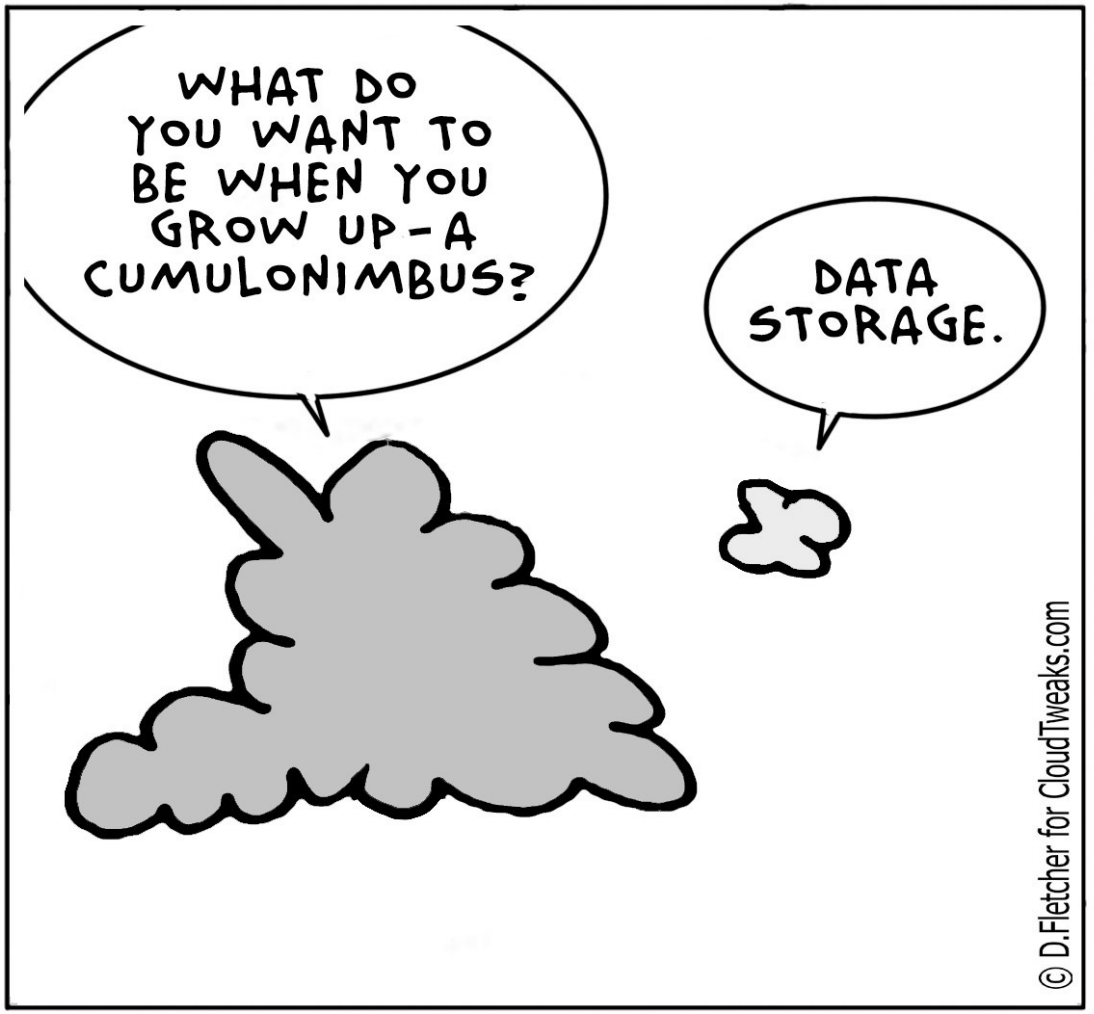
Encryption on server side

- Partially lost confidentiality for clients
  (server has access to decrypted data)

Data at-rest – combination of ...

- Full disk encryption
- Filesystem encryption
- Object store encryption

Questions?


WHAT DO YOU WANT TO BE WHEN YOU GROW UP - A CUMULONIMBUS?
DATA STORAGE.
© D.Fletcher for CloudTweaks.com