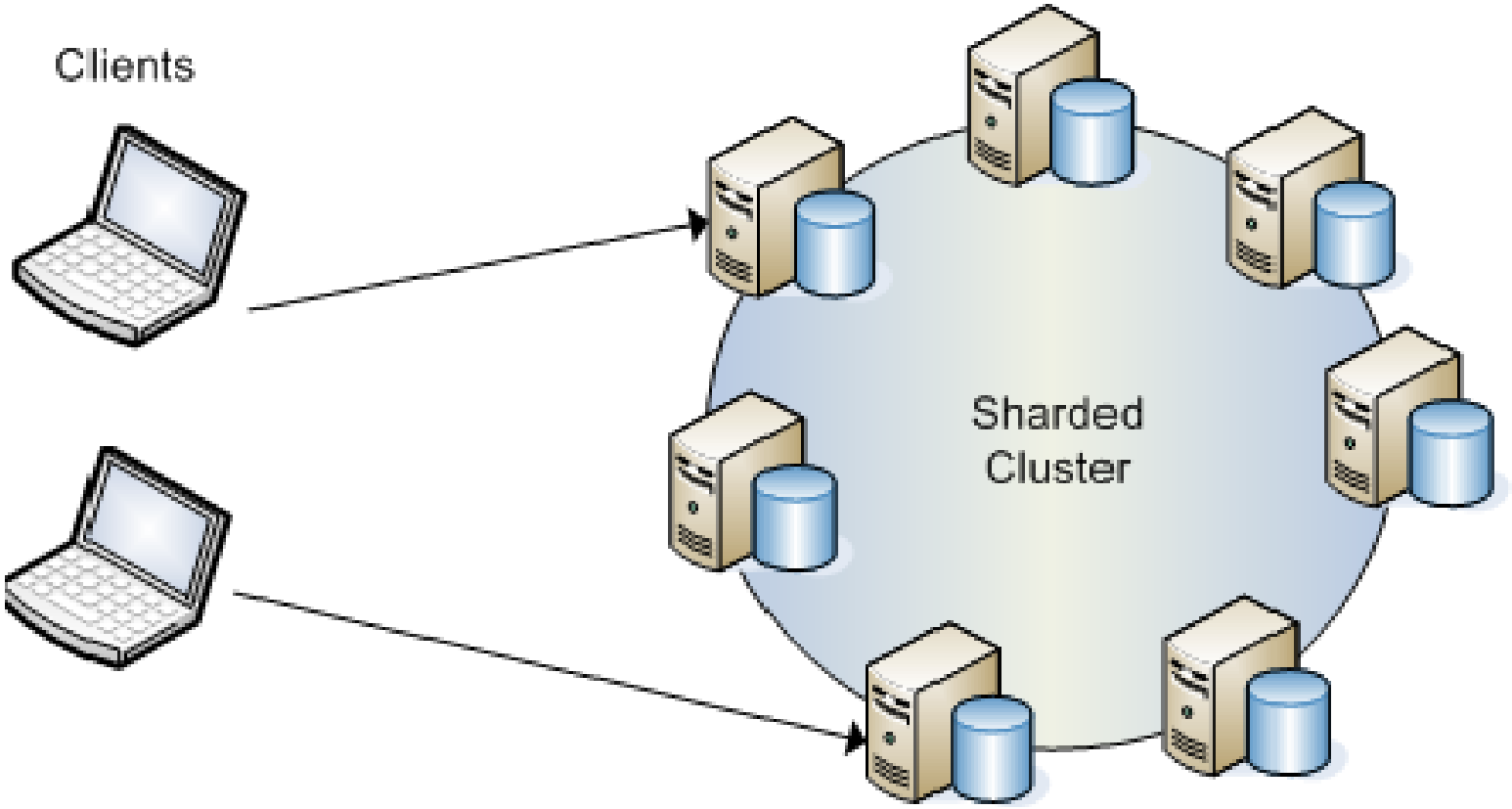
A Ceph, mint adattároló klaszter megoldás

készítette: Szalai László,
Major Kálmán (NYME INGA)

Elosztott adattárolás séma



Mik merülhetnek fel

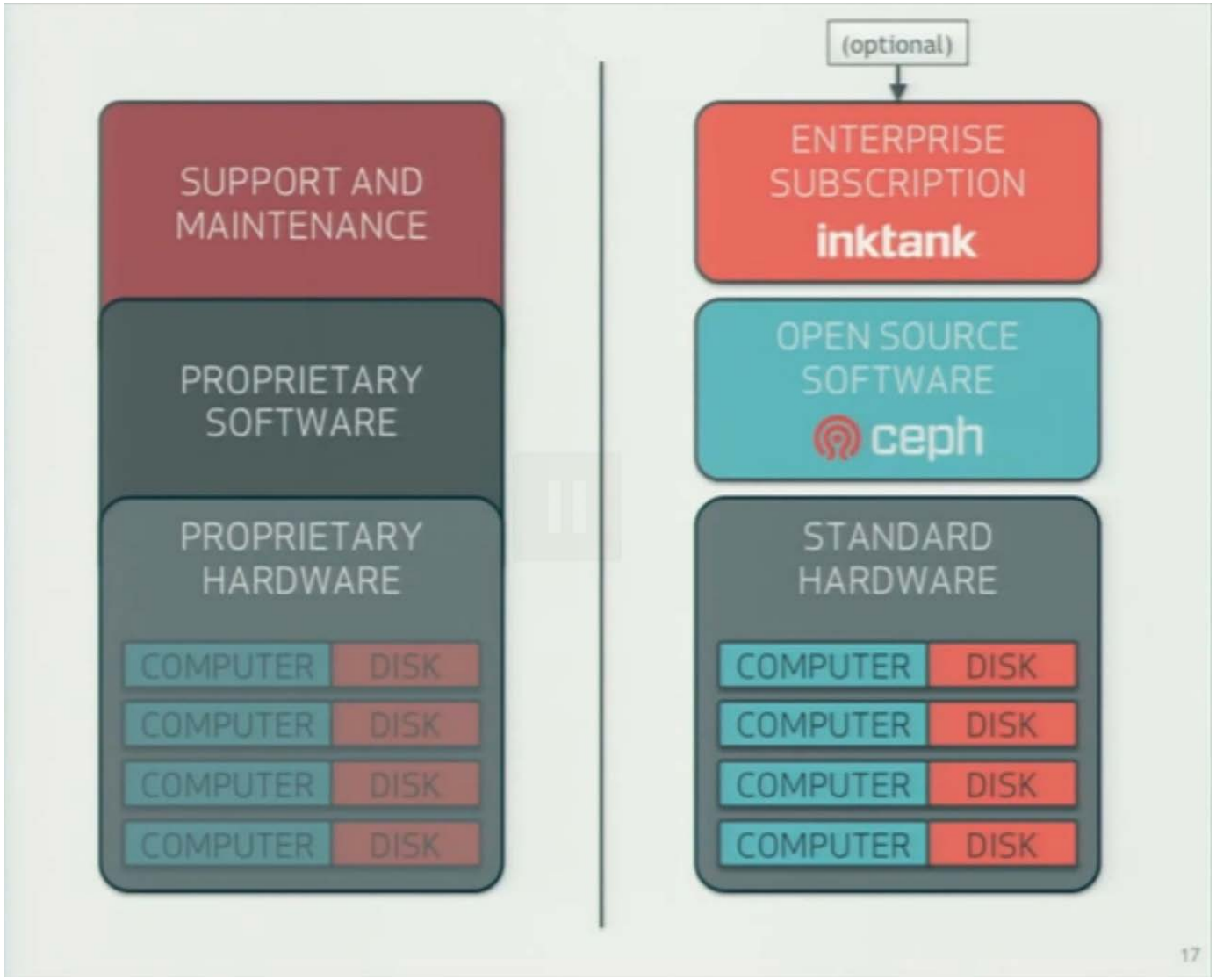
If you need...	Gluster	Lustre	Ceph
NAS and Scale Out NAS	✔	✔	✔
SAN			✔
Shared Filesystems	✔	✔	✔
Object Storage	✔		✔

(MooseFS, XtreemFS, stb.)

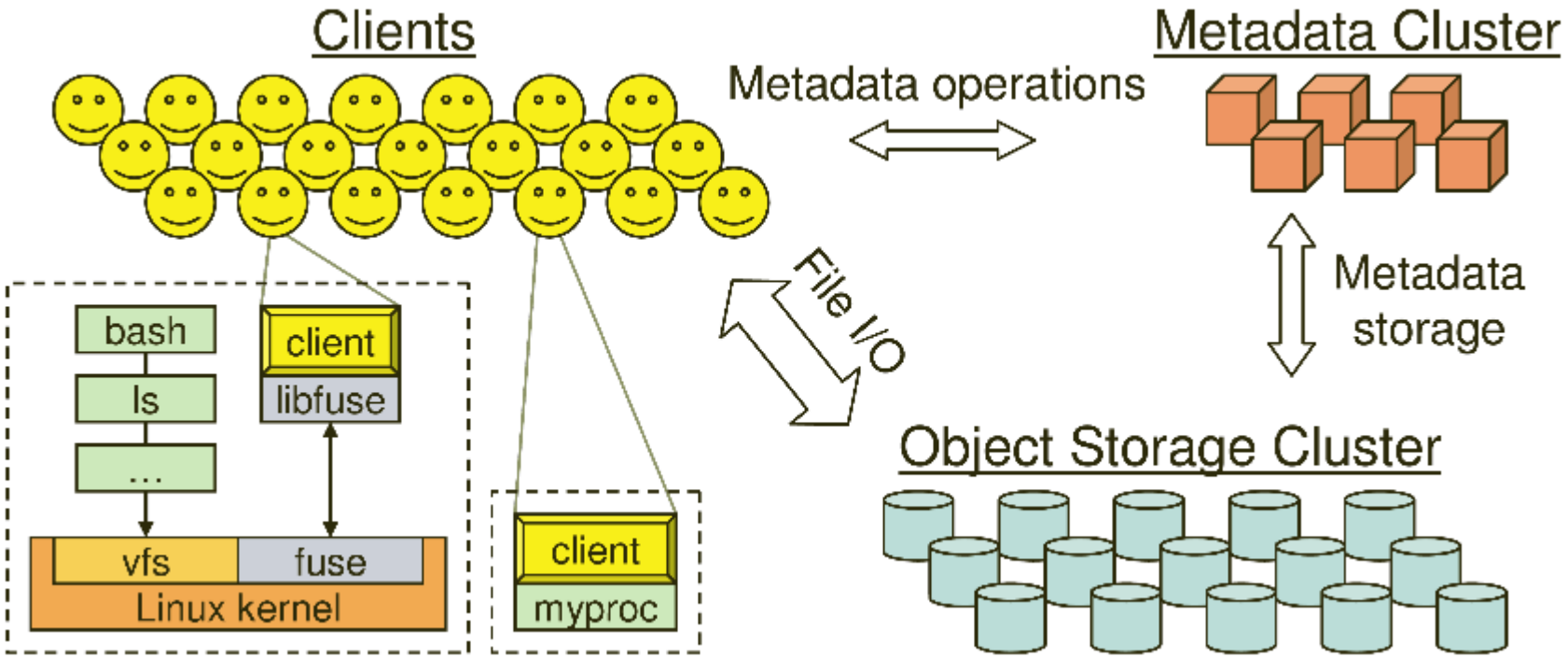
ceph, mint választott

- Antik görög kifejezés
- kezdetben doktori munka
- 2007-től három fős fejlesztőbrigád
- 2012, Inktank Storage cég
  - support

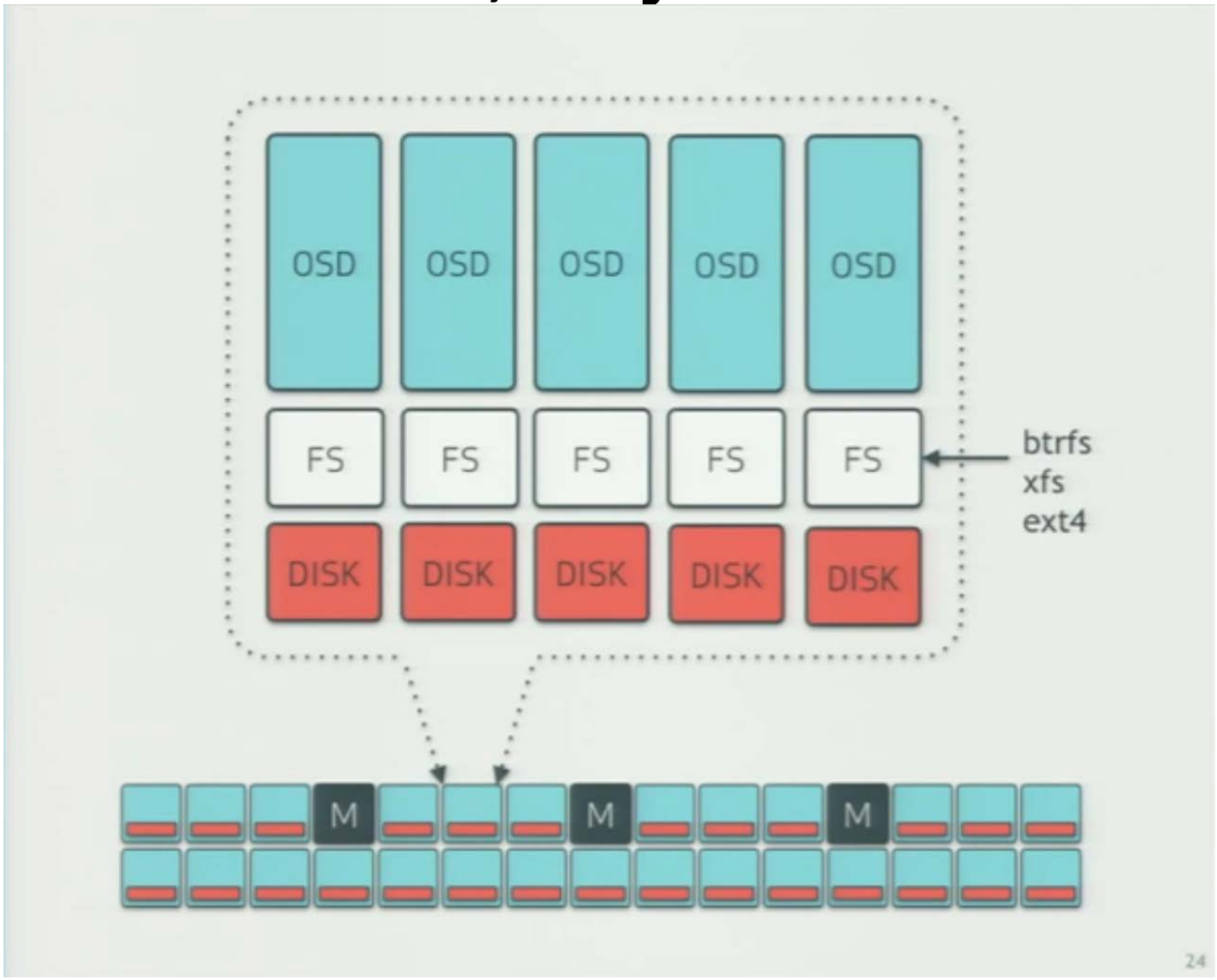
Filozófia


SUPPORT AND MAINTENANCE
PROPRIETARY SOFTWARE
PROPRIETARY HARDWARE
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
(optional)
ENTERPRISE SUBSCRIPTION
inktank
OPEN SOURCE SOFTWARE
ceph
STANDARD HARDWARE
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
17

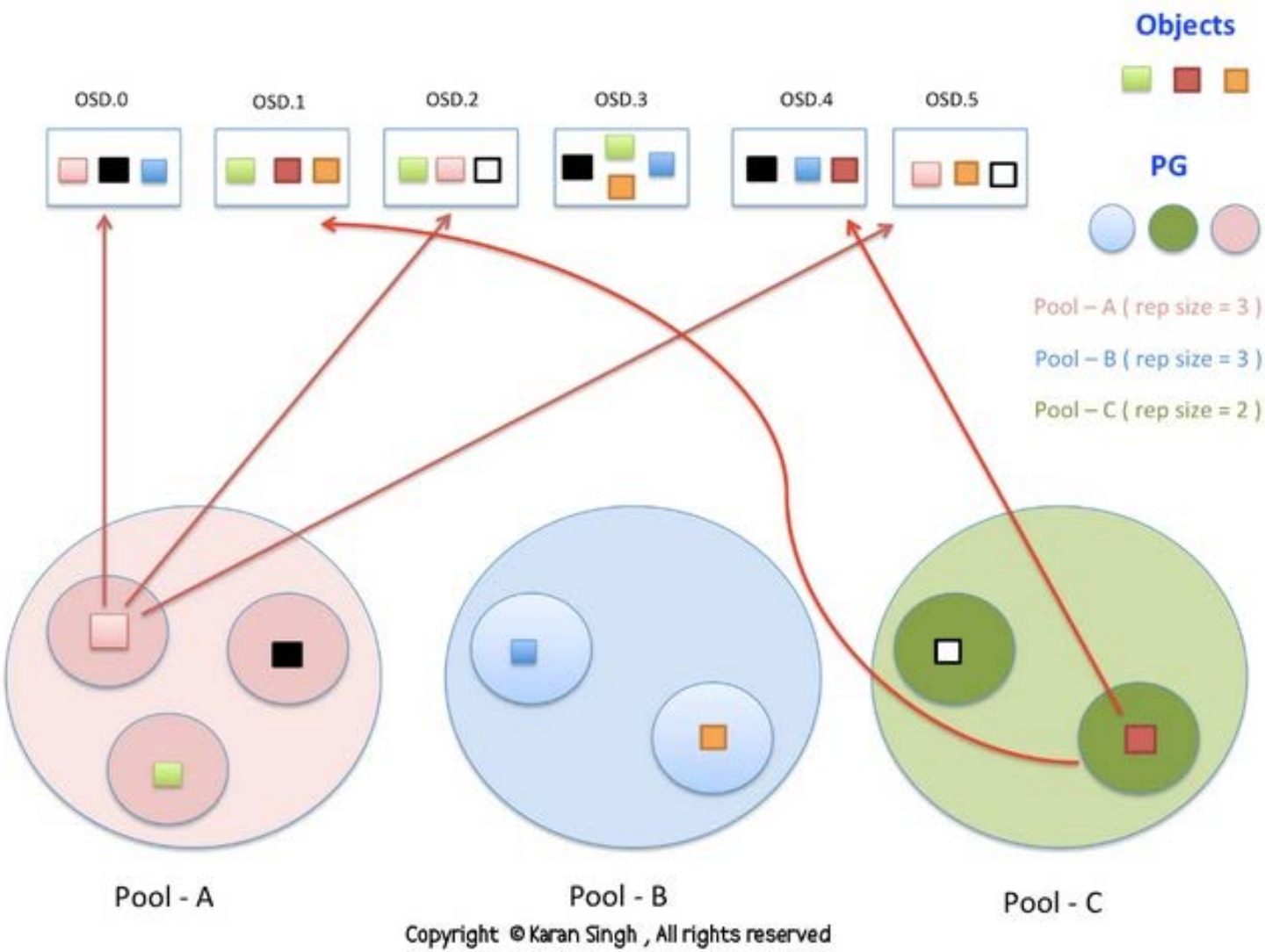
Hogyan, miként?



Adattárolás, objektumszintű



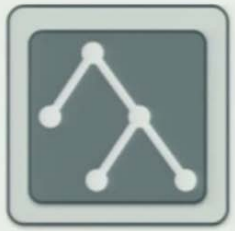
Hol az adat?


Objects
PG
OSD.0
OSD.1
OSD.2
OSD.3
OSD.4
OSD.5
Pool – A (rep size = 3)
Pool – B (rep size = 3)
Pool – C (rep size = 2)
Pool - A
Pool - B
Pool - C
Copyright © Karan Singh , All rights reserved

Felépítő kockák

Monitors:
- Maintain cluster map
- Provide consensus for distributed decision-making
- Must have an odd number
- These do **not** serve stored objects to clients

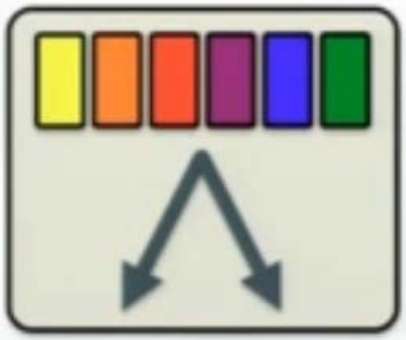
OSDs:
- One per disk (recommended)
- At least three in a cluster
- Serve stored objects to clients
- Intelligently peer to perform replication tasks
- Supports object classes



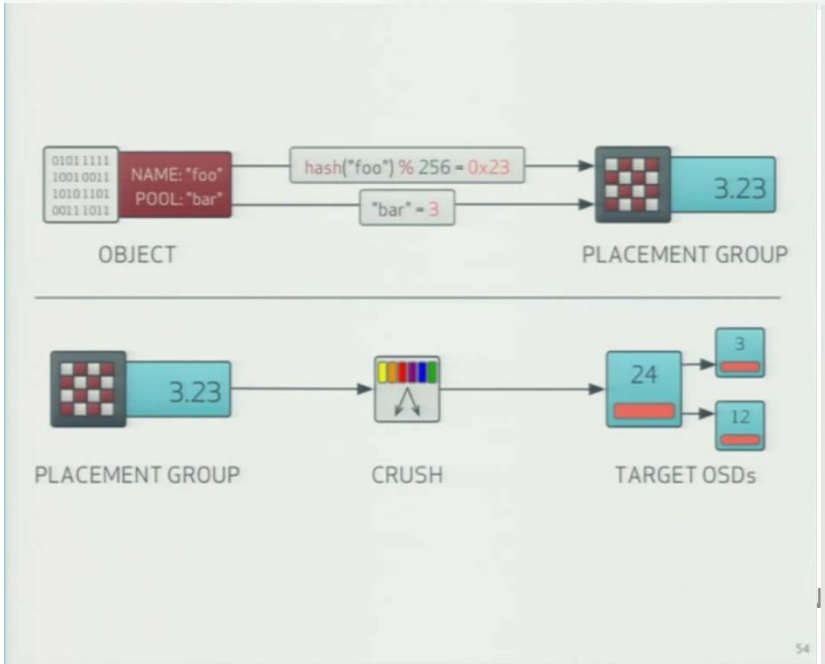
Metadata Server
- Manages metadata for a POSIX-compliant shared filesystem
  - Directory hierarchy
  - File metadata (owner, timestamps, mode, etc.)
- Stores metadata in RADOS
- Does not serve file data to clients
- Only required for shared filesystem

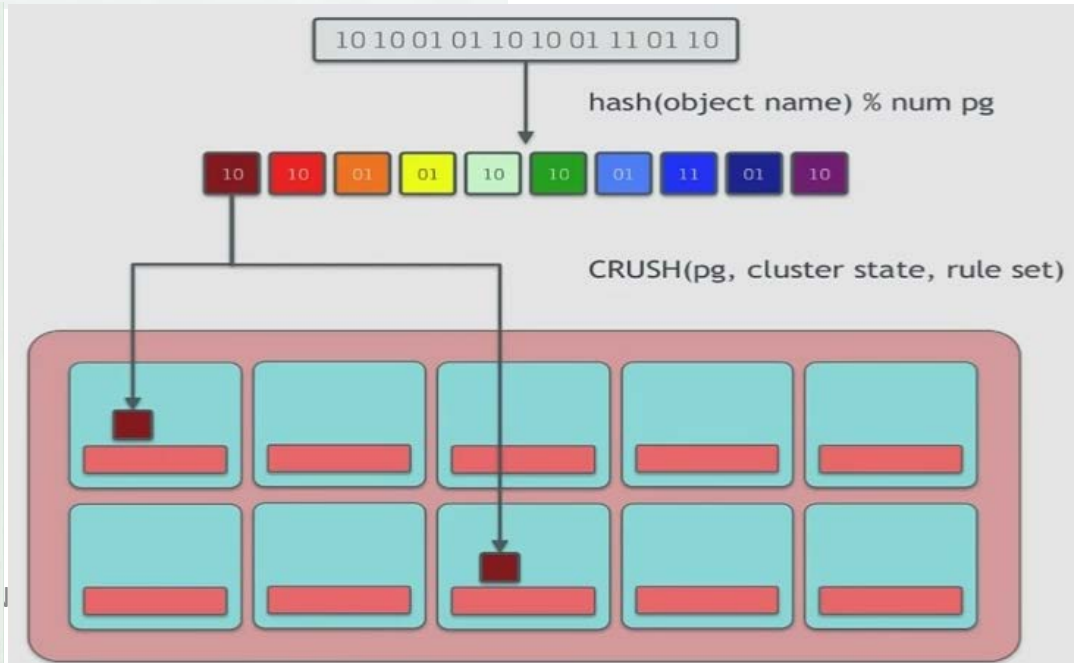
40

CRUSH algoritmus

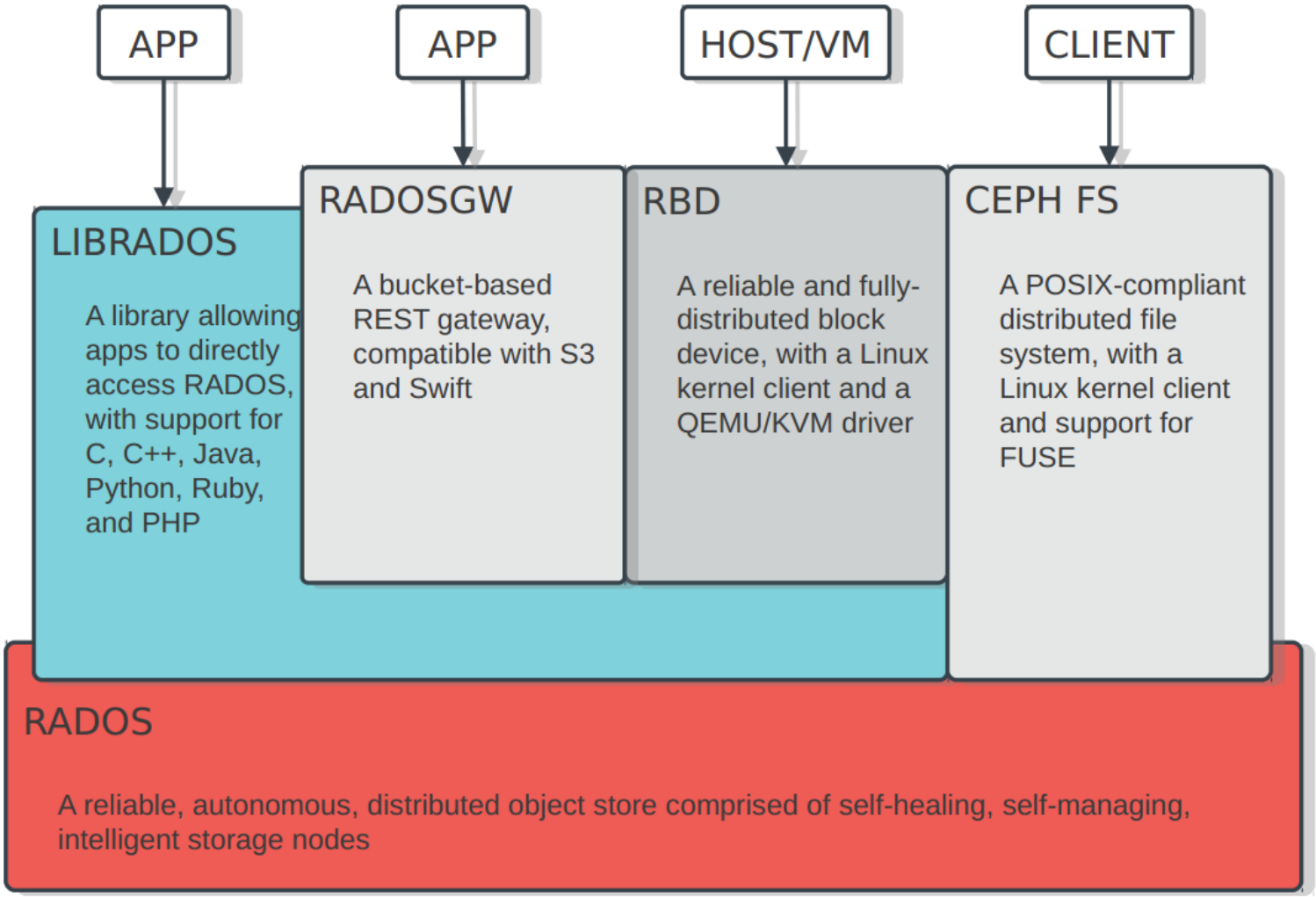


- CRUSH
  - Pseudo-random placement algorithm
  - Ensures even distribution
  - Repeatable, deterministic
  - Rule-based configuration
    - Replica count
    - Infrastructure topology
    - Weighting





Tudáslista

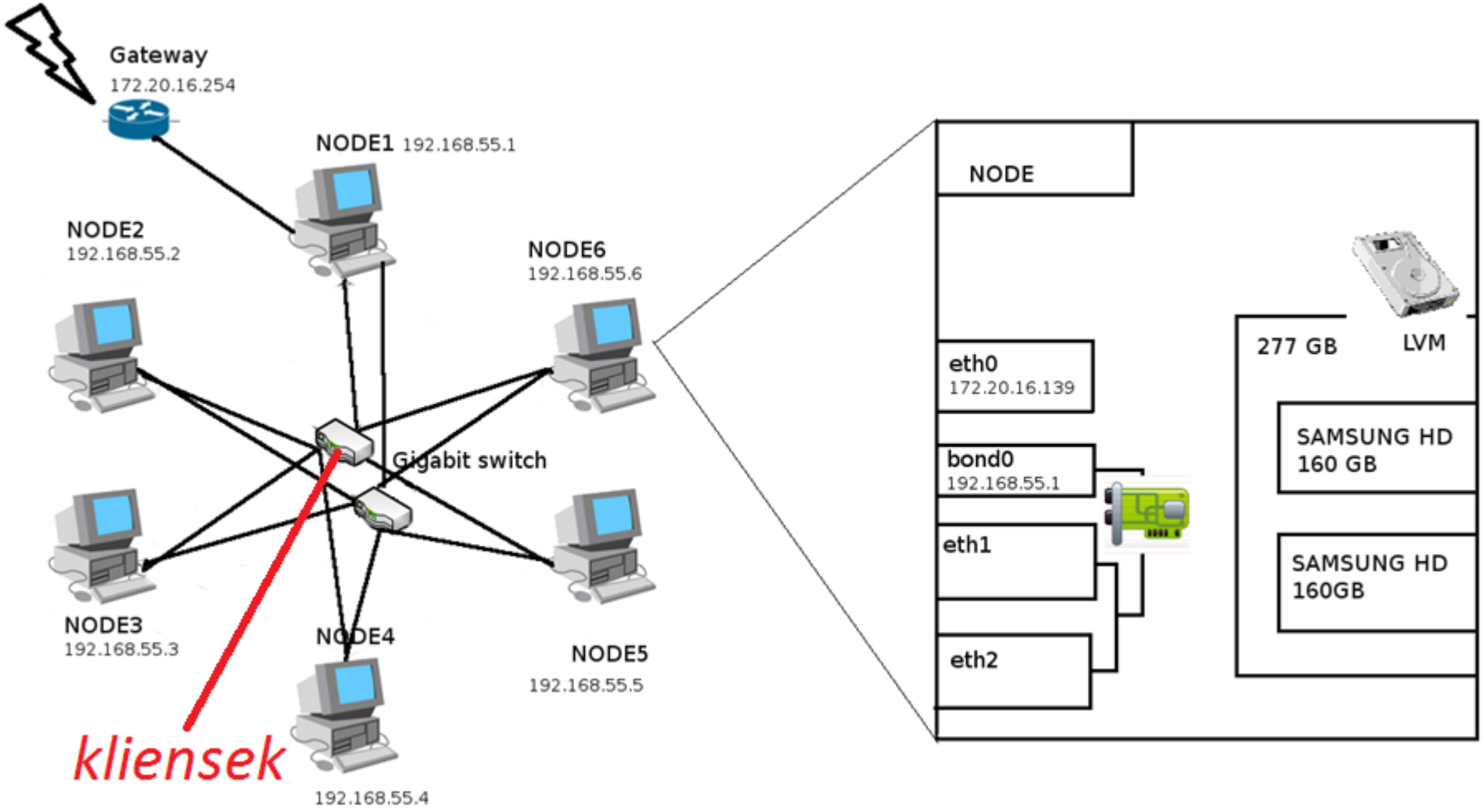

APP
APP
HOST/VM
CLIENT
LIBRADOS
A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP
RADOSGW
A bucket-based REST gateway, compatible with S3 and Swift
RBD
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver
CEPH FS
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE
RADOS
A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes

Hogy is kezdtük #1

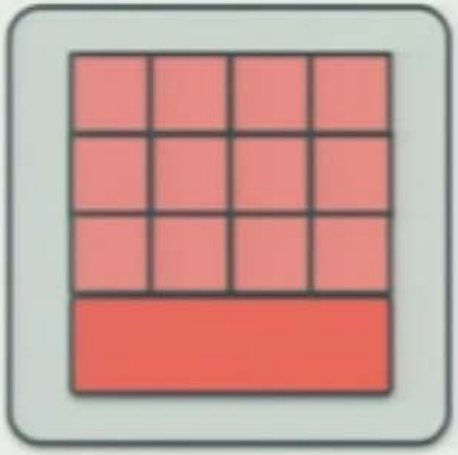
- Régebbi Core2Duo asztali gépek
- 2 GB RAM
- 2 x 160 GB SATA2 HDD
- 2 Gigabit NIC

- 2 db Gigabit Office Switch, CAT5 UTP kábelek

Hogy is kezdtük #2, topológia



Blokkeszköz!



RADOS Block Device:

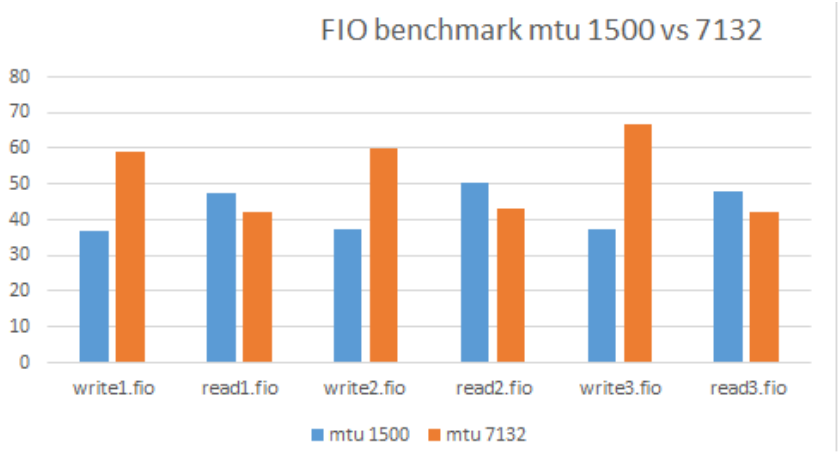
- Storage of disk images in RADOS
- Decouples VMs from host
- Images are striped across the cluster (pool)
- Snapshots
- Copy-on-write clones
- Support in:
  - Mainline Linux Kernel (2.6.39+)
  - Qemu/KVM, native Xen coming soon
  - OpenStack, CloudStack, Nebula, Proxmox

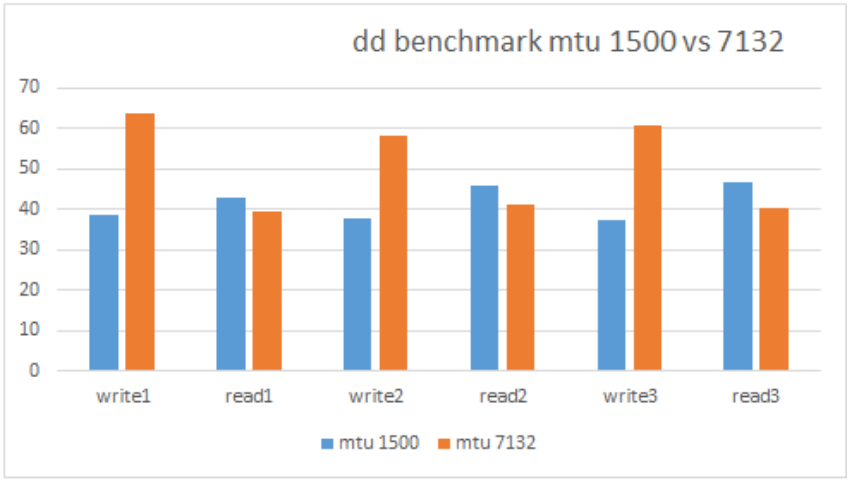
És filerendszer!

- Komplett metadata tárolás
- POSIX „kompatibilitás”
- NFS szerű csatolás és működés
- Jól működő flock() és fcntl() kezelés

- Nem ajánlott egyelőre produktív üzemben

Sebességtesztek





Egyéb tesztek

- Failover
  - RBD ✓
  - CephFS ✓
- Cluster startup, recovery ✓
- Snapshot (brtfs) ?
- Live migration ?
- Performance tuning !