

Ceph Storage

- Ceph ist ein redundanter, über viele Server verteilter Objektspeicher
- Bietet verschiedene Arten von Speicher:
  - S3-/Swift-API-kompatibler Objektspeicher (RADOS Gateway)
  - Virtuelle Block Storage (RADOS Block Device / RBD)
  - Verteiltes Cluster File System (CephFS)
  - Alles im gleichen Cluster
- Ab 3 Server bis zu tausenden von Servern skalierbar

© 2023 Marco Gabriel, inett GmbH Proxmox Server Solutions GmbH



Ceph Features

- Software Defined Storage, Open Source
- Läuft auf Standardhardware, kein RAID Controller
- Object Storage, Block Storage, File System
- Snapshots, Copy-On-Write, Cloning
- Thin Provisioning, Selbstheilend, Skalierbar bis Exabyte Level
- Resizable Images, Import/Export von Images
- Client direkt in Linux und QEMU KVM integriert
- Caching und Auto-Tiering
- Pools für HDD, SSD, NVMe
- CRUSH Algorithmus (Controlled Replication Under Scalable Hashing)
- Nahezu unlimitierte Storage
- NFS, iSCSI Gateways verfügbar
- Kommerzieller Support von Proxmox, Red Hat, Suse, Ubuntu und anderen IT Service Providern

© 2023 Marco Gabriel, inett GmbH Proxmox Server Solutions GmbH

ceph

CERN @CERN

Today's #PhotoOfTheWeek takes us to the CERN Data Centre.

It can store and analyse more than 600 petabytes of data (600 million gigabytes) generated by the collisions during #BeamTime of #LHCRun3.

It corresponds to over 20 000 years of 24/7 HD video recording. #ComputingatCERN

Tweet übersetzen



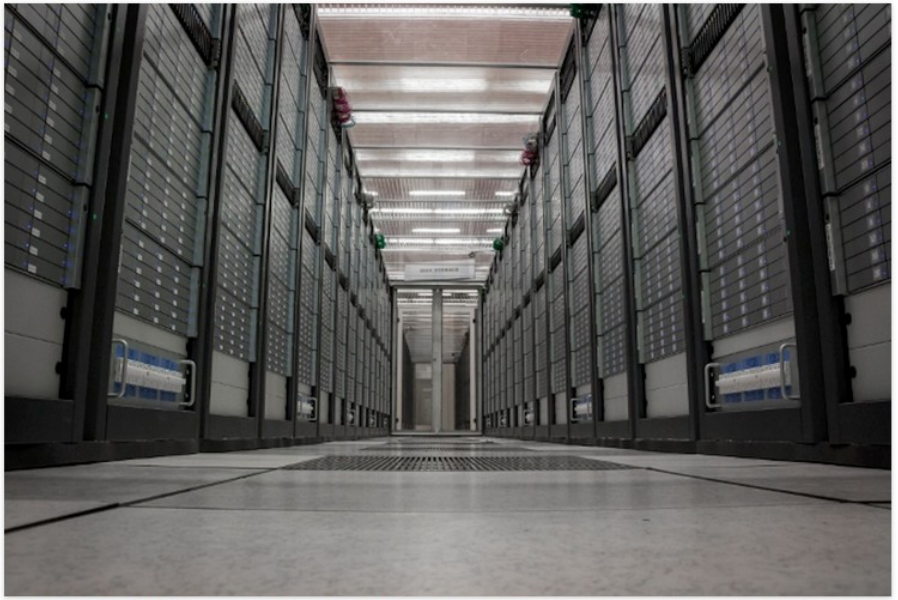
5:03 nachm. · 5. Apr. 2023 · **63.300** Mal angezeigt

https://twitter.com/cern/status/1643630413682823170?s=46&t=ZU8FN9LZT46tI1pPVUdoYw

An exabyte of disk storage at CERN

CERN disk storage capacity passes the threshold of one million terabytes of disk space

29 SEPTEMBER, 2023 | By Tim Smith



A fraction of the 111 000 devices that form CERN's data storage capacity. (Image: CERN)

CERN's data store has now crossed the remarkable capacity threshold of one exabyte, meaning that CERN has one million terabytes of disk space ready for data!

https://home.cern/news/news/computing/exabyte-disk-storage-cern

© 2023 Marco Gabriel, inett GmbH Proxmox Server Solutions GmbH



Ceph in Proxmox VE

- Proxmox VE kann entweder
  - An externe Ceph Cluster angebunden werden
  - Als Storage Komponente für einen hyperkonvergenten Betrieb von Proxmox VE (HCI) verwendet werden
- Im Falle von Hyperkonvergenz (HCI)
  - Ceph Administration und Management über die Proxmox GUI möglich
  - Ceph Dashboard in der Proxmox GUI vorhanden
  - Ceph CLI Befehle können genutzt werden

© 2023 Marco Gabriel, inett GmbH Proxmox Server Solutions GmbH



Voraussetzungen

- 10Gbit+ Netzwerk für Ceph Backend
- Mindestens 4 Disks pro Server
- Ähnliche Disks und Server in Clustern < ~10 Nodes
- Ceph benötigt RAM und CPU
  - Minimum pro Storage Device (OSD):
    - ~5 GB RAM
    - 1 GHz, 1 Core
  - Auch mgr, mon und andere Komponenten benötigen noch zusätzlichen Speicher und Rechenleistung

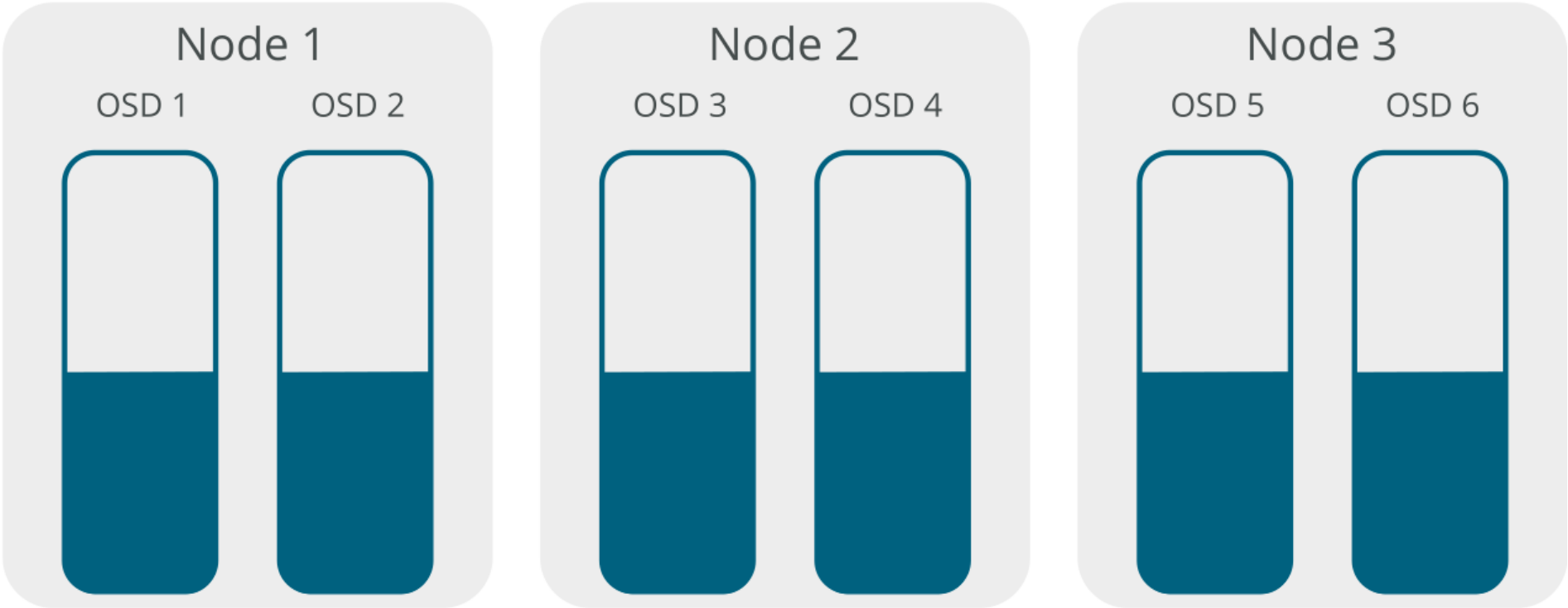
© 2023 Marco Gabriel, inett GmbH Proxmox Server Solutions GmbH

Ceph Pool Types

- Replicated
  - Speichert ***copies / replicas*** auf mehreren Hosts (Redundanz)
- Erasure Code (EC)
  - Splittet Daten in ***k*** Chunks und erzeugt ***m*** Coding Chunks. Jeder Chunk wird auf einem separaten Server gespeichert.

Too few OSDs per node

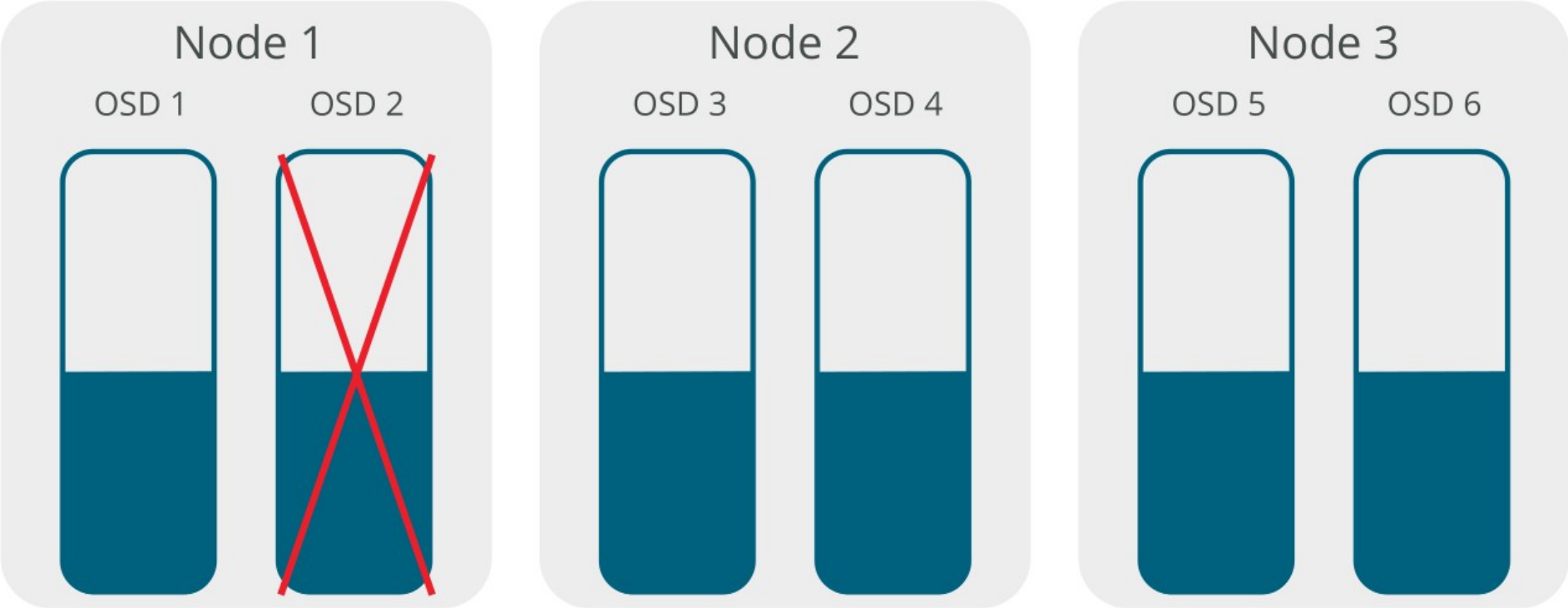
Normal Operation



© 2023 Marco Gabriel, inett GmbH

Too few OSDs per node

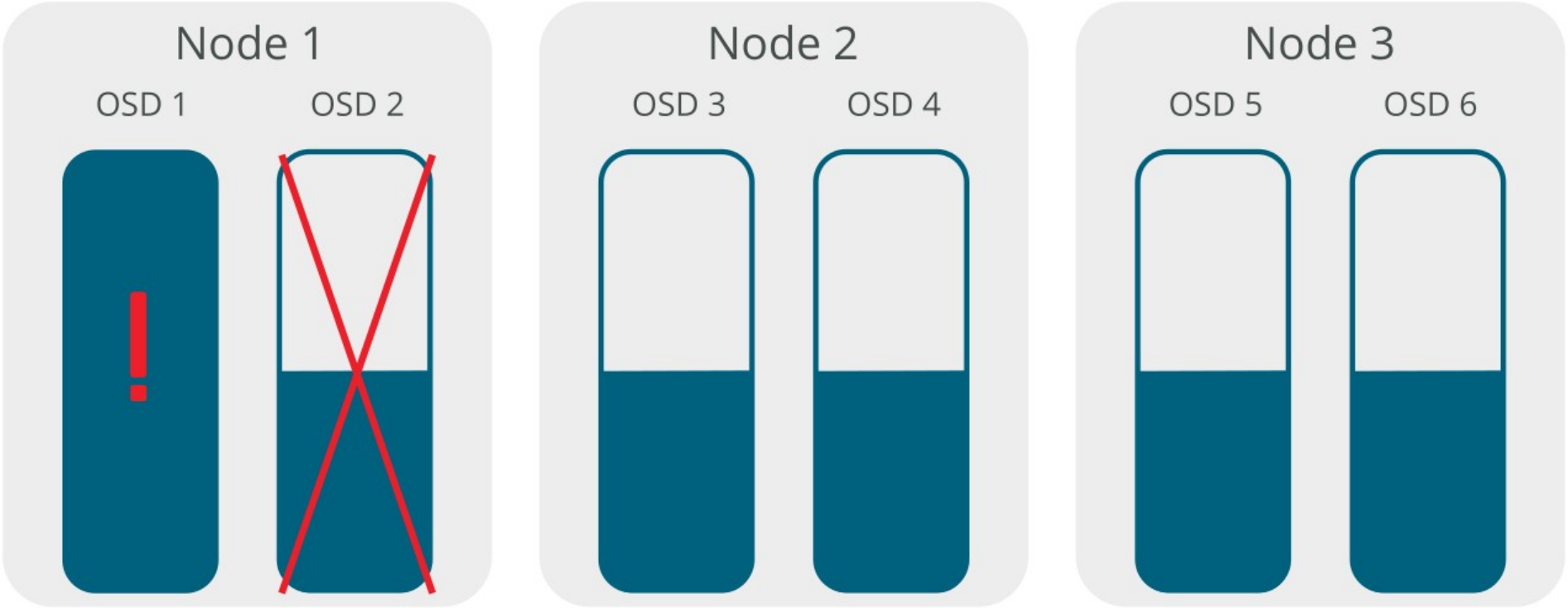
Device failure



© 2023 Marco Gabriel, inett GmbH Proxmox Server Solutions GmbH

Too few OSDs per node

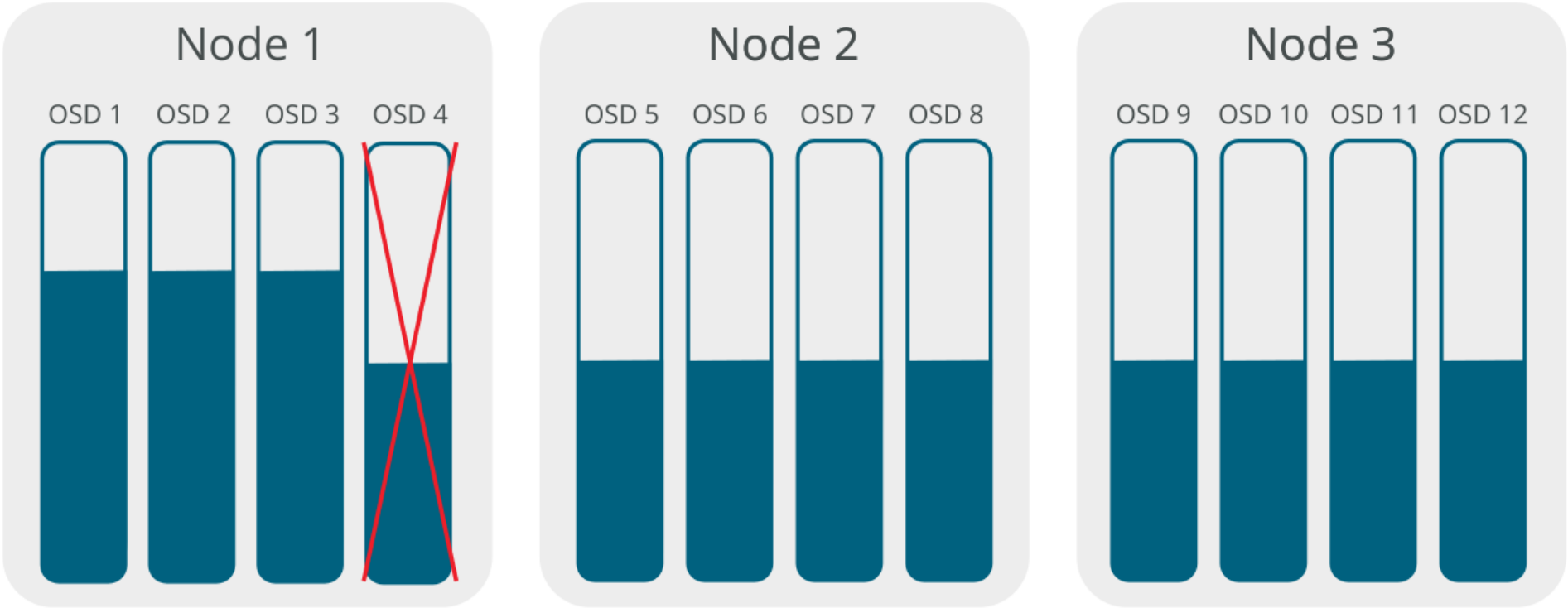
Working device recovers until full
then pool stops working



© 2023 Marco Gabriel, inett GmbH

Too few OSDs per node

Easy to avoid by using at least 4 or even more devices per server



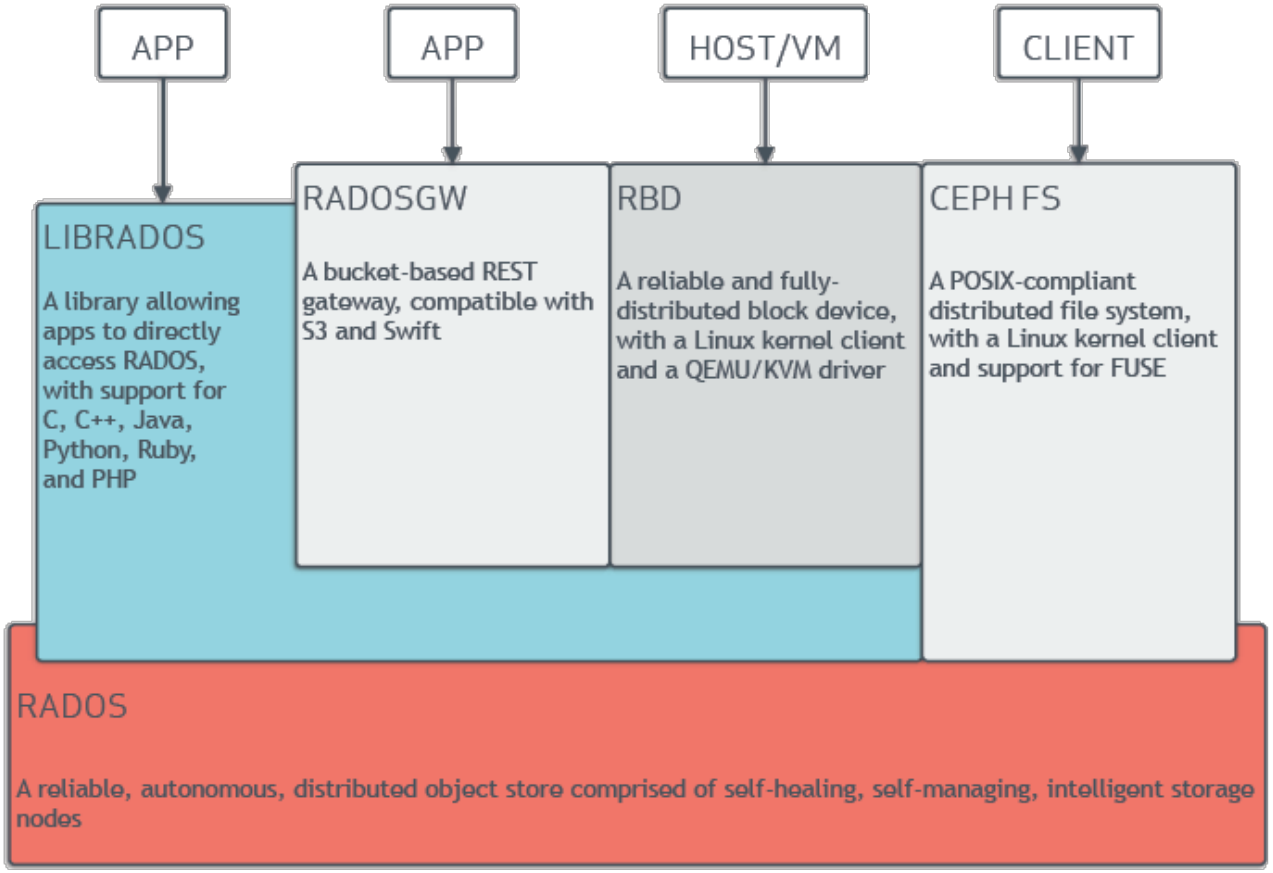
© 2023 Marco Gabriel, inett GmbH

Kapazitätsplanung

- Standard: Replicated Pool (size=3, min_size=2)
- Beispiel:
  - Benötigt für VMs: 10 TB
  - „Festplattenverschnitt“: 10 TB * 1,1 = 11 TB
  - 3-fache Replikation: 11 TB * 3 = 33 TB
  - Max. Space Usage < 85% = 33 TB / 0.85 = 38.82 TB
  - Objekte sind nicht absolut gleich verteilt, Abweichung ~10%: 38.82 TB / 0.9 = 43.13 TB
  - Wenn ein Server ausfällt werden Objekte auf den restlichen Nodes wiederhergestellt → Mehr Speicher nötig: 1 von 5 fällt aus = 20% → 43.13 TB / 0.8 = 54 TB
  - 54 TB auf 5 Servern = ~11 TB pro Server
  - 11 TB = 4x 3.68 TB SSDs/NVMe pro server

© 2023 Marco Gabriel, inett GmbH Proxmox Server Solutions GmbH

Ceph Architektur



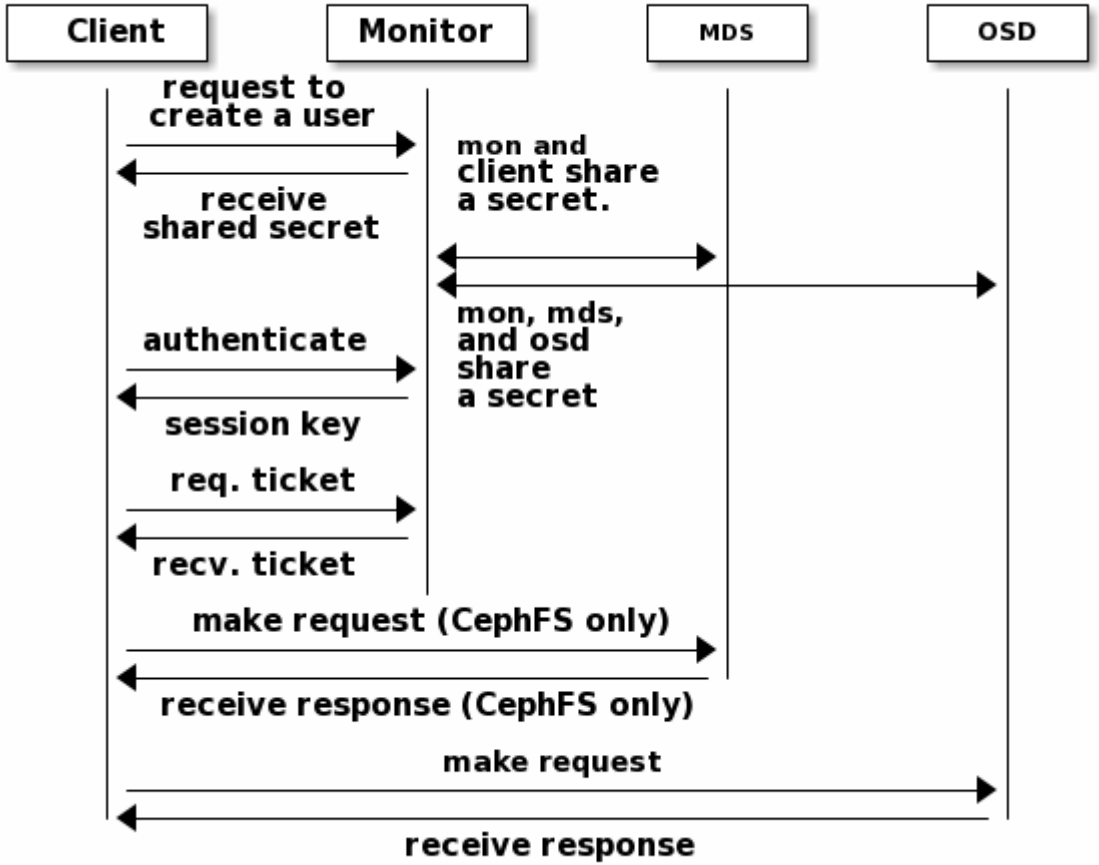
© 2023 Marco Gabriel, inett GmbH Proxmox Server Solutions GmbH

Ceph Komponenten



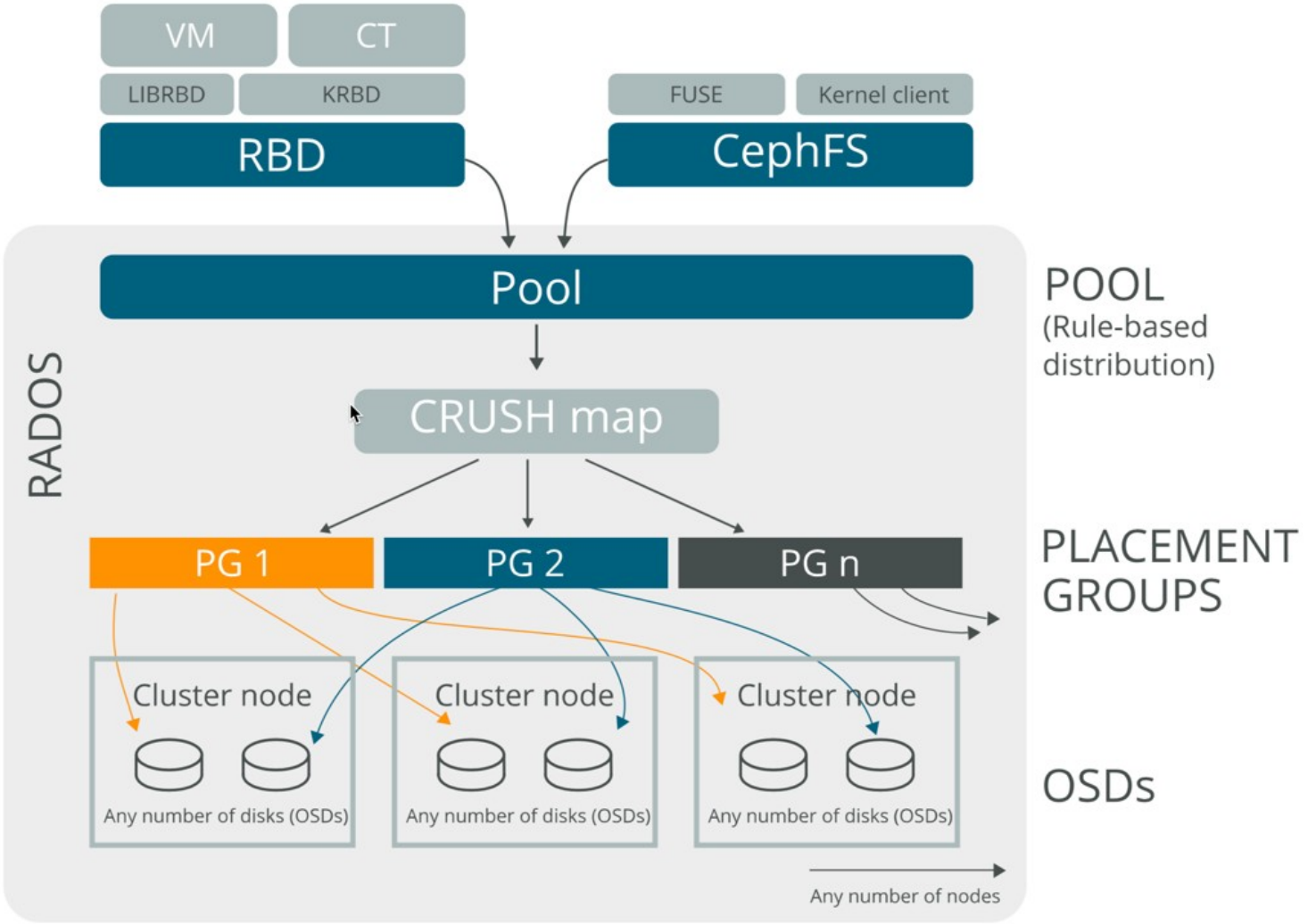
© 2023 Marco Gabriel, inett GmbH

Ceph Requests



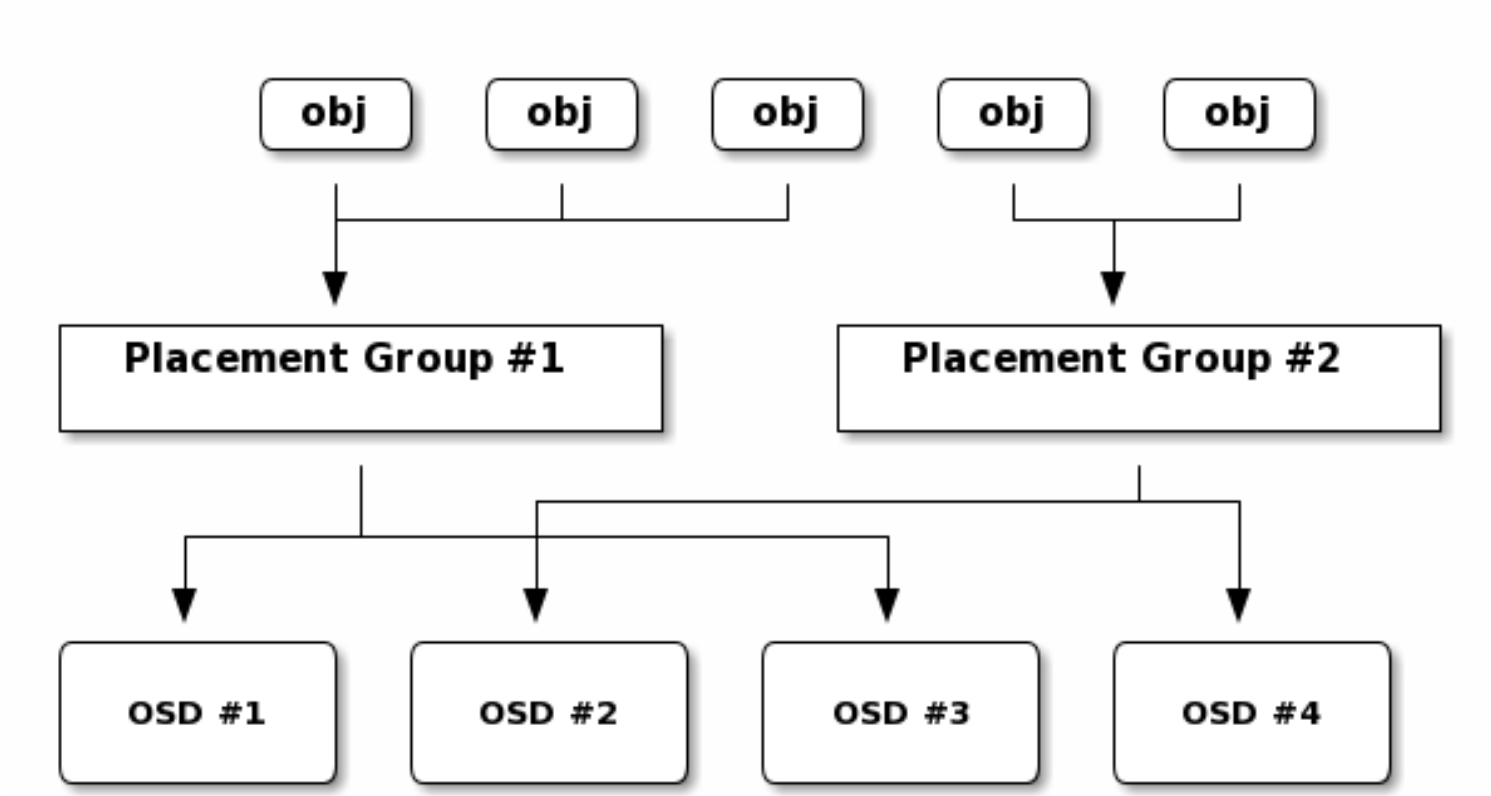
© 2023 Marco Gabriel, inett GmbH

Ceph



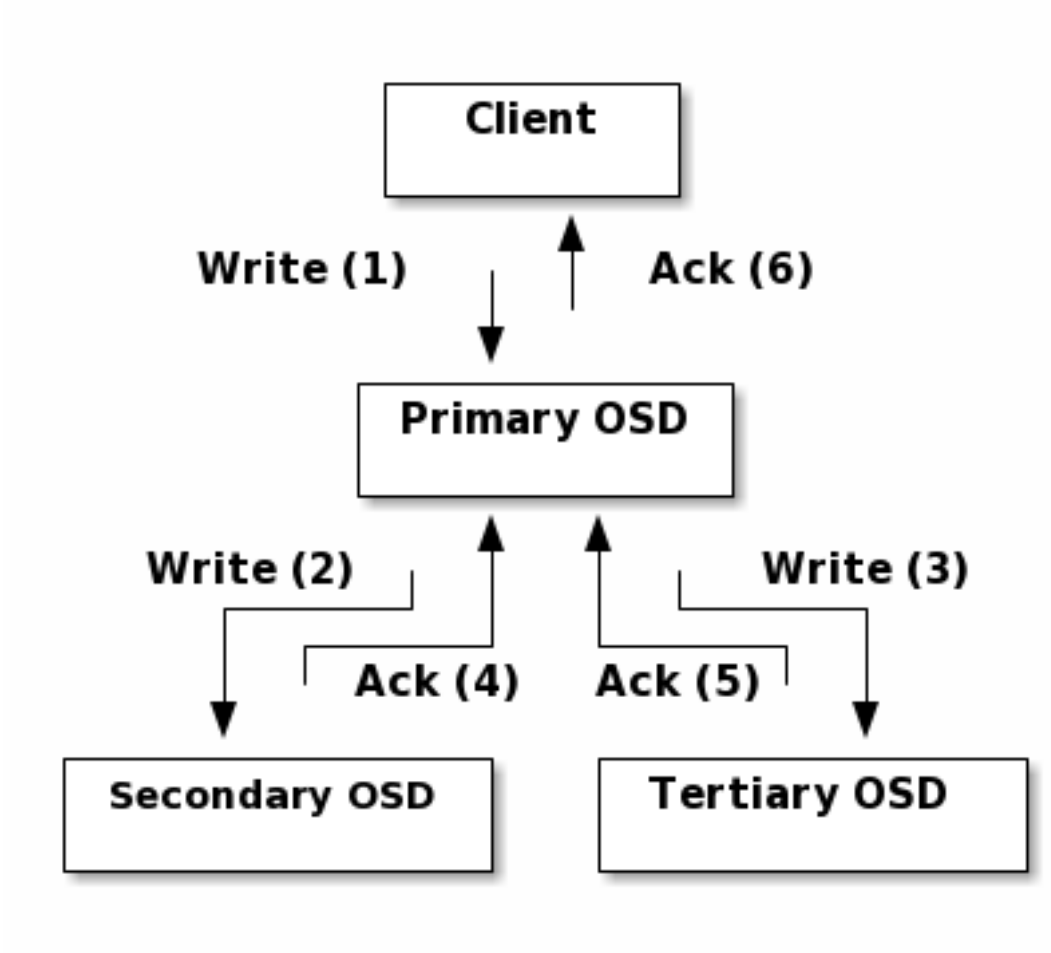
© 2023 Marco Gabriel, inett GmbH

Ceph Placement Groups (PGs)



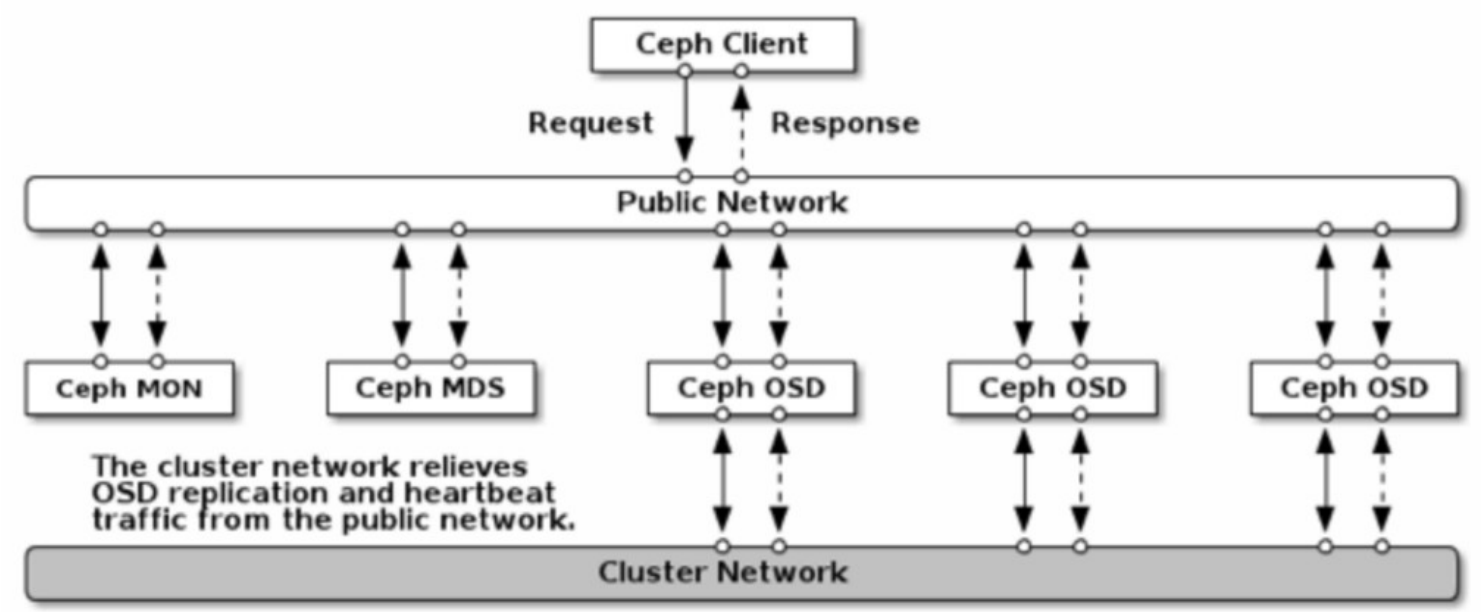
© 2023 Marco Gabriel, inett GmbH

Ceph Schreibvorgang



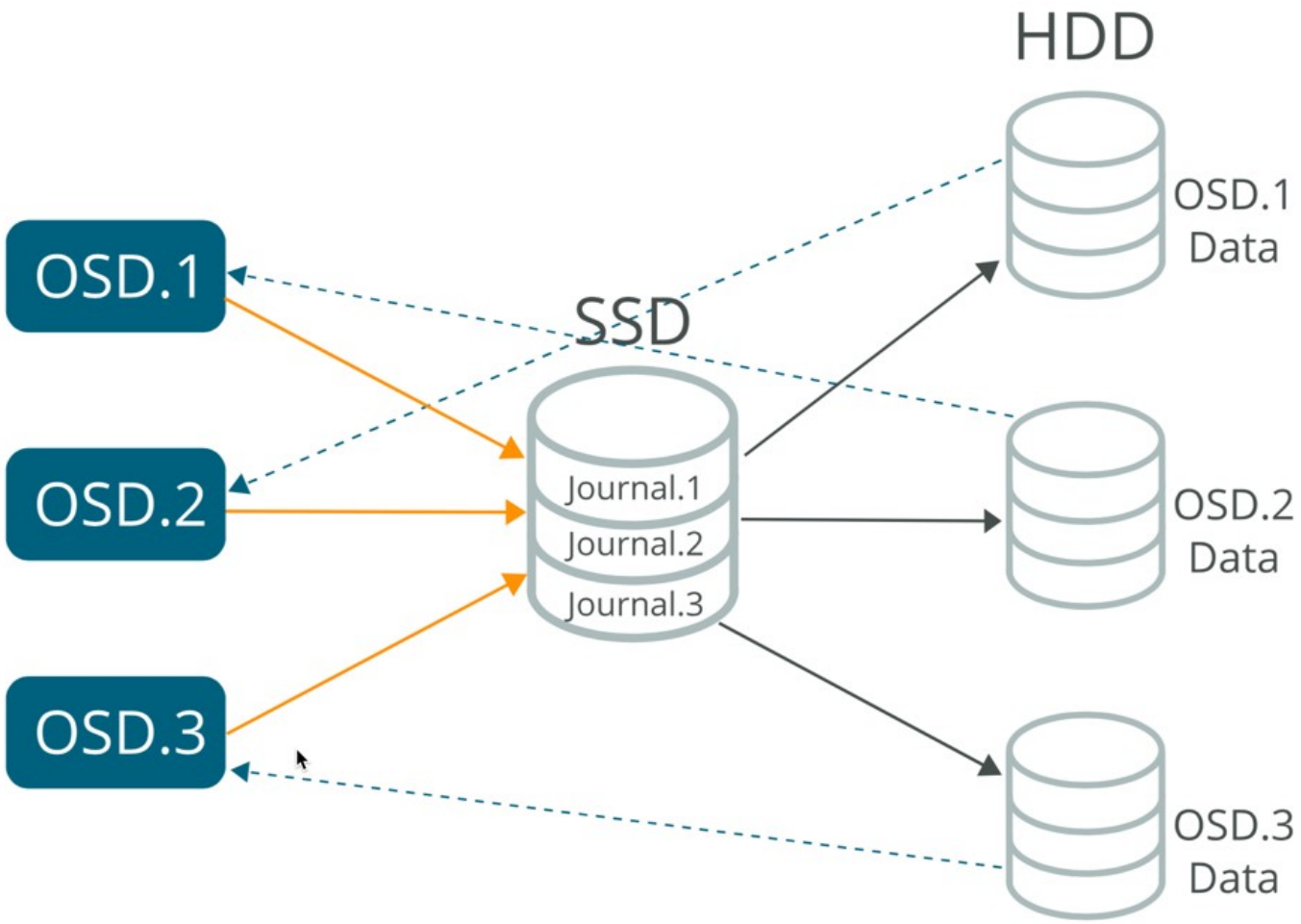
© 2023 Marco Gabriel, inett GmbH Proxmox Server Solutions GmbH

Ceph Netzwerk(e)



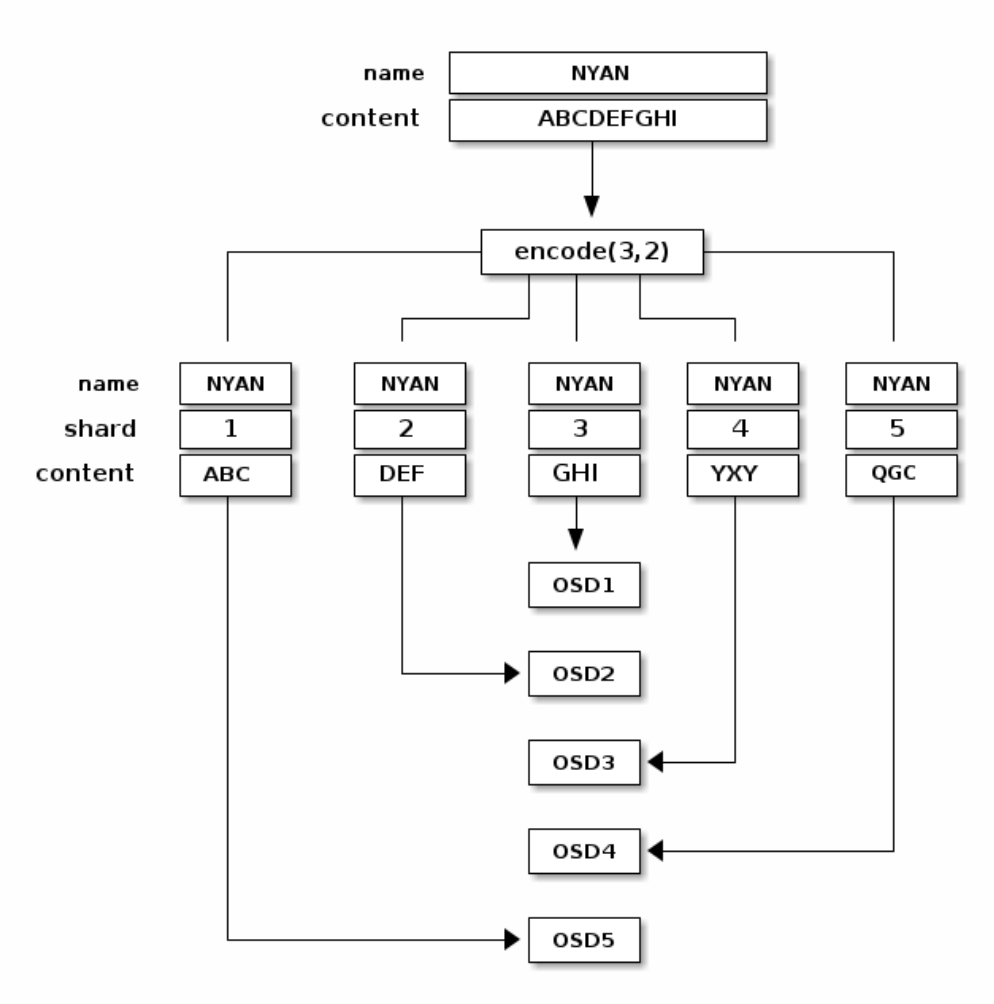
© 2023 Marco Gabriel, inett GmbH

Ceph Journal / SSD Journal

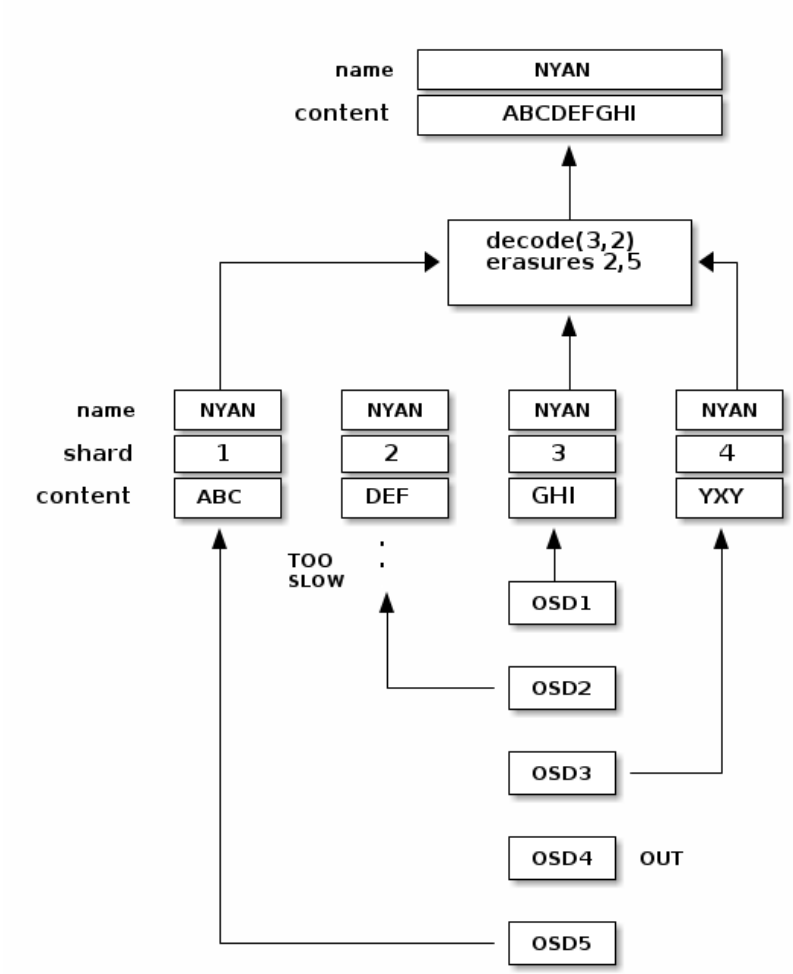


© 2023 Marco Gabriel, inett GmbH

Erasure Coding - Schreibvorgang



Erasure Coding - Lesevorgang



Die wichtigsten Befehle

- `ceph status` zeigt den Status / Zustand des Clusters allgemein an
- `ceph health` zeigt nur die Gesundheit an
- `ceph health detail` zeigt die Gesundheit detailliert an (und gibt evtl. Hinweise zur Fehlerbehebung)