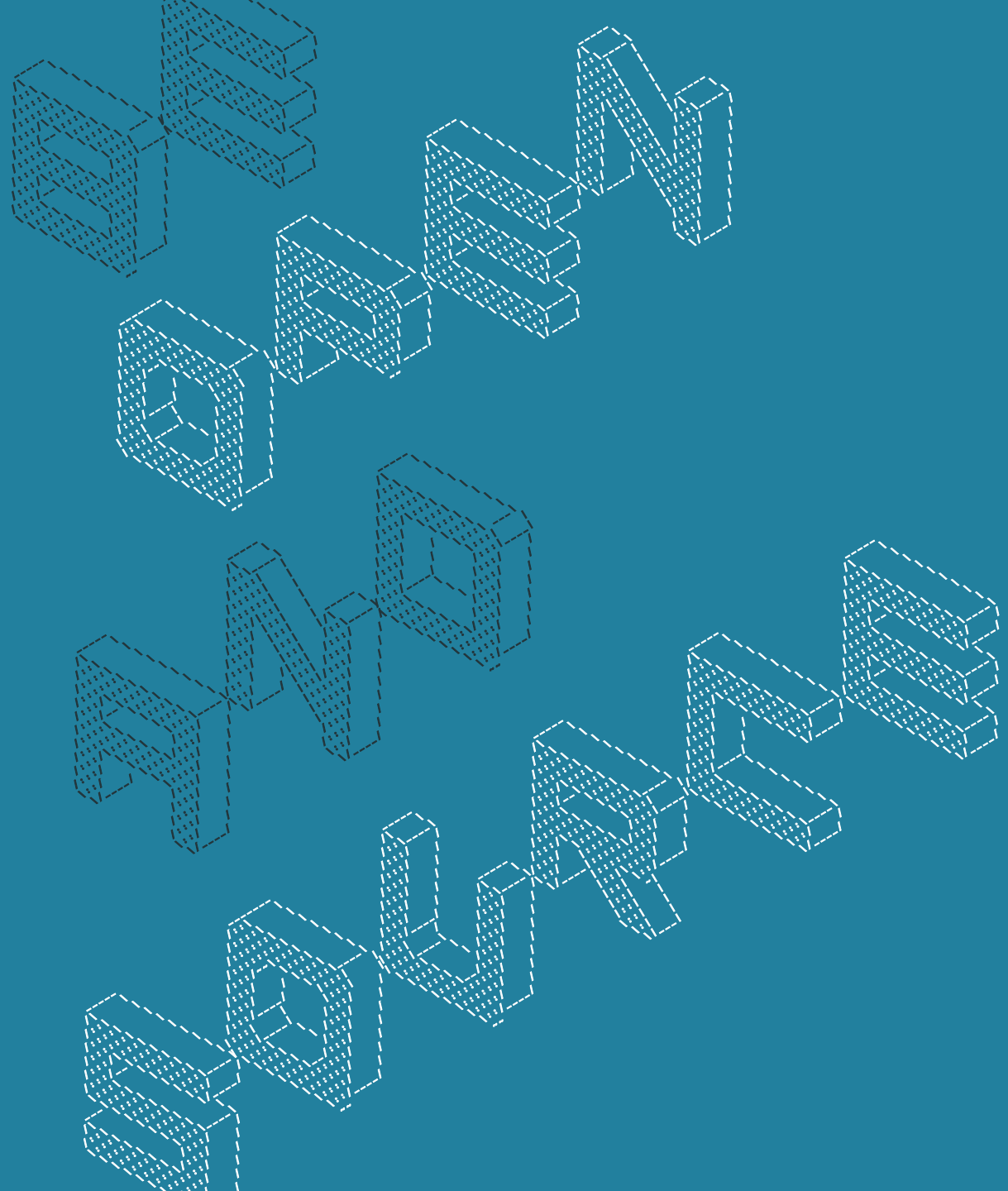


SLAC 2026

11.–13. Mai 2026 | Berlin

www.slac-2026.de

Einführung und Neues in Ceph

Ceph

Ceph ist …

gesetzt

- gibt es seit 2006
- Doktorarbeit von Sage Weil

interessant

- verteilter Objektspeicher
- Redundanz
- Datensicherheit
- effiziente Skalierung
- lauffähig auf (fast) jeder Hardware

Ceph bietet …

einen Storage-Cluster

- der sich selbst verwaltet
- der sich selbst heilt
- ohne Engpässe

drei Schnittstellen

- Objektspeicher (kompatibel zu S3)
- Blockspeicher (für VMs, etc)
- verteiltes Dateisystem

Object Storage Daemon

OSD speichert Daten auf

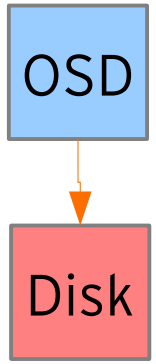
- HDD
- SSD
- NVMe
- oder was es noch geben wird

Ein Prozess pro Blockdevice

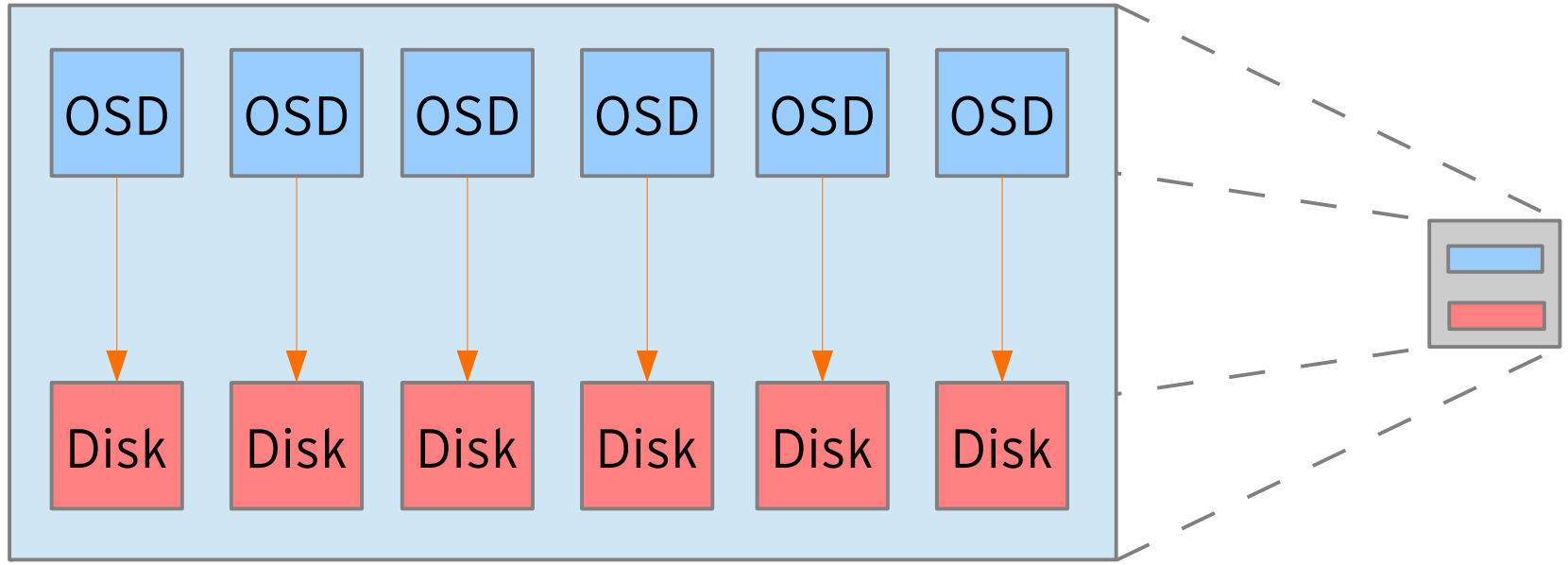
OSDs liefern Daten direkt an Clients

OSDs sprechen mit anderen OSDs

- Replikation
- Datenwiederherstellung



Mehrere OSDs in einem Ceph-Knoten



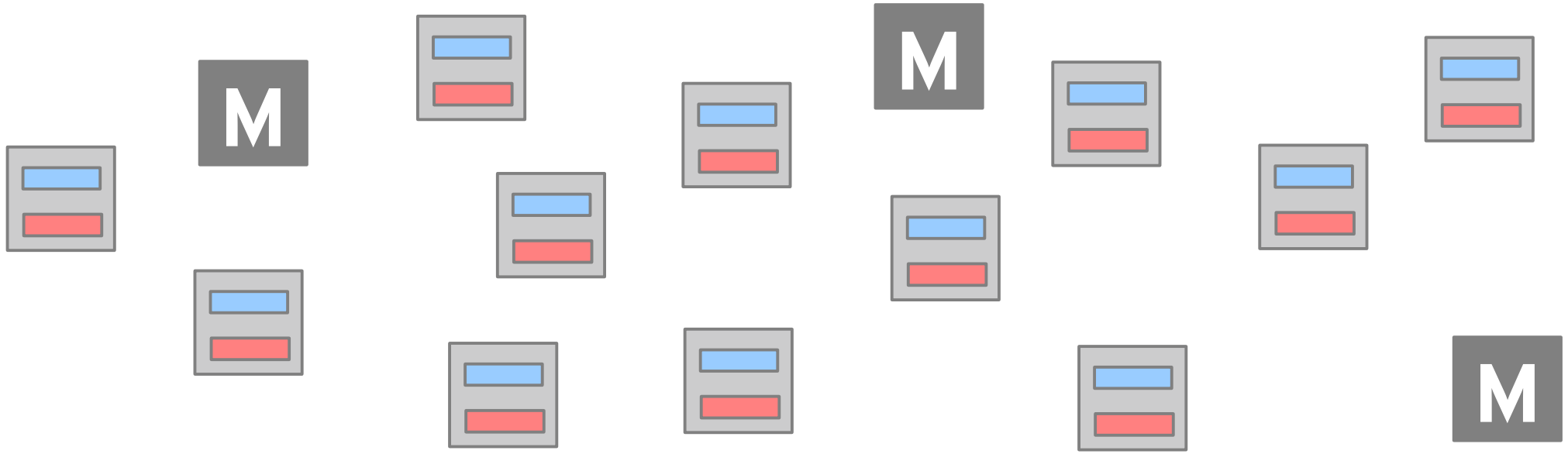
Dazu kommen Monitore

MONs bilden das Gehirn des Clusters

Quorum für Entscheidungen

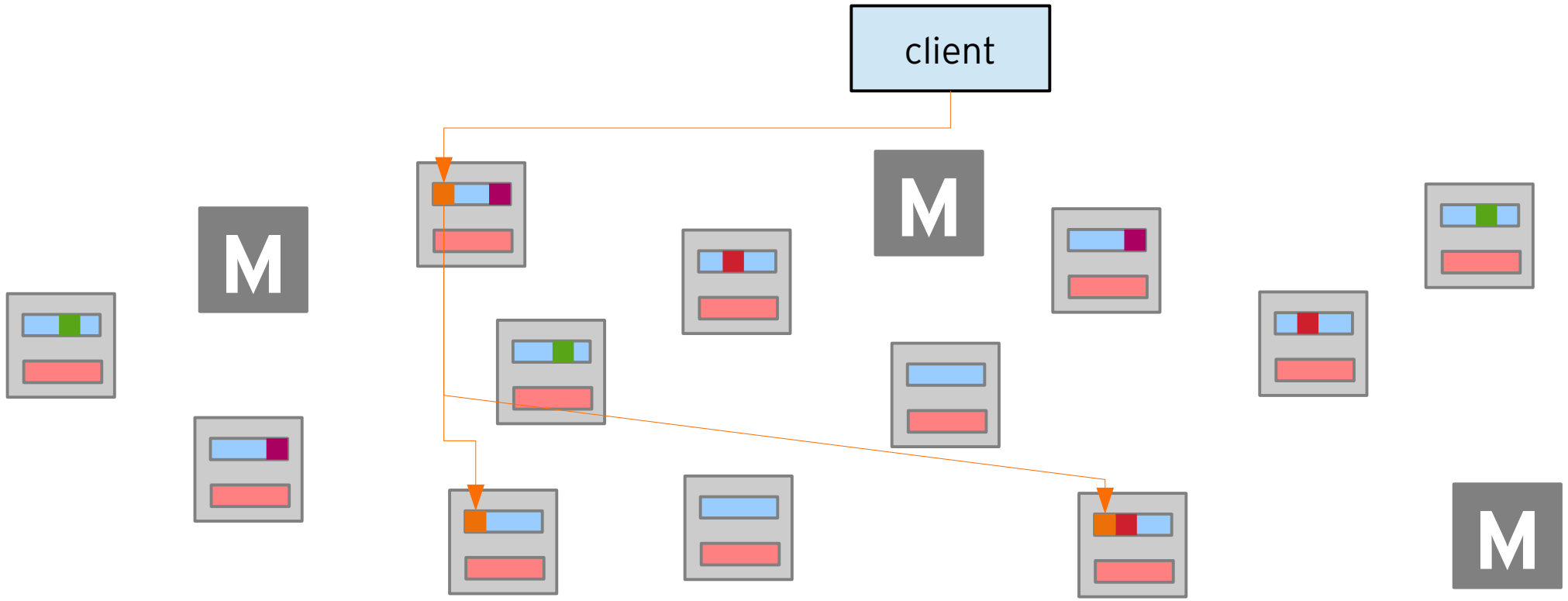
Nicht im Datenpfad

Und schon ist der Ceph-Cluster fertig



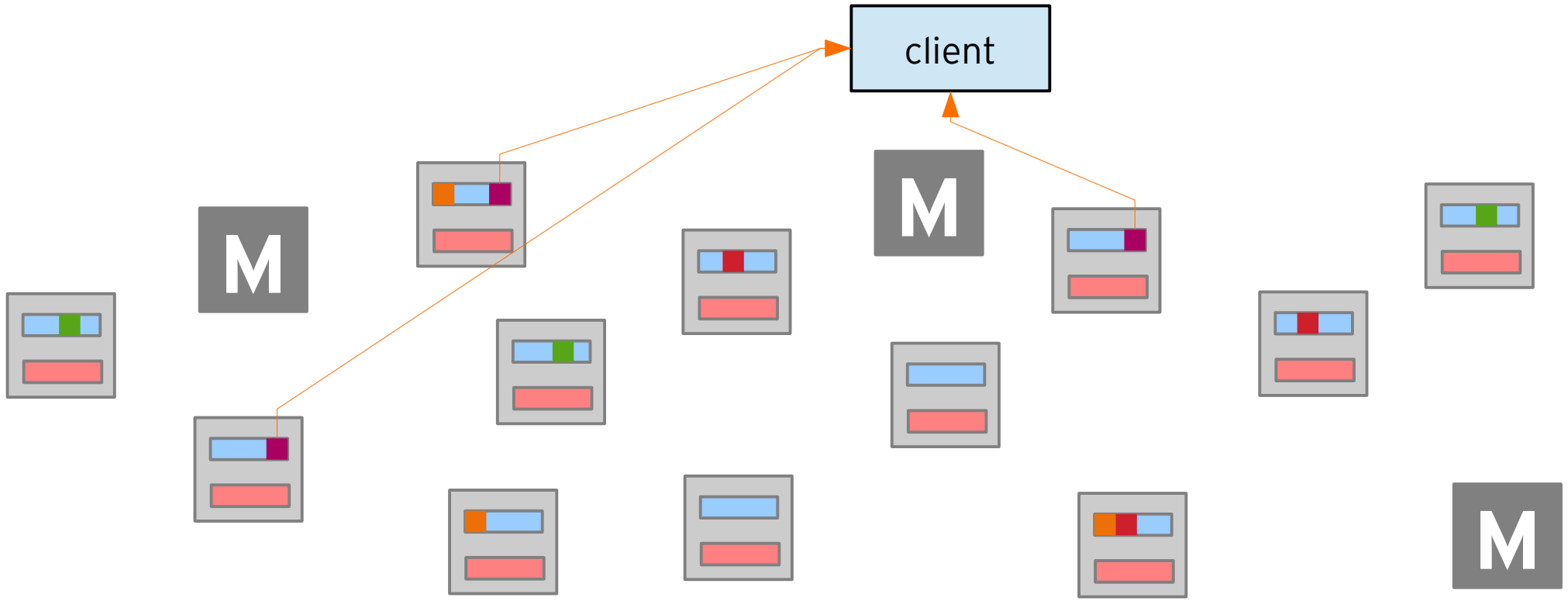
Schreiben …

Ceph-Objekt wird auf ein OSD geschrieben und von dort repliziert



… und Lesen

gelesen werden kann von jeder Kopie



Konzeptionelle Komponenten

Pool

logischer Container für Ceph-Objekte

Eigenschaften

- Name + ID
- Anzahl der Objektkopien
- Erasure Coding Einstellungen
- CRUSH Regel
- Besitzer
- Quota
- Snapshots

CRUSH

Controlled Replication Under Scalable Hashing

- Algorithmus

MONs verwalten CRUSH Map

- Topologie des Clusters
- Ausfallzonen

Clients kalkurieren selber

- Platzierung der Daten
- keine Engpässe im Datenpfad

Redundanz

Replikation

n genaue Kopien

hohe Leseraten

gute Schreibraten

- schnelles Cluster-Netzwerk

Wiederherstellung von mehreren Quellen

Netto-Kapazität nur 1/n

- n = 3 → 33%

n + 1 unabhängige Knoten notwendig

Erasure Coding

Daten aufgeteilt in k Teile und m Parity

Platzeffizient

hoher CPU-Verbrauch beim Schreiben

- Parity-Berechnung

Wiederherstellung braucht CPU

Netto-Kapazität: k / (k + m)

- k = 8, m = 2 → 80%
- k = 2, m = 2 → 50%

k + m + 1 unabhängige Knoten notwendig

Ceph-Clients

librados

native Ceph-Objekte

wird praktisch nicht verwendet

`rados` als CLI-Tool praktisch für Debugging etc

RADOS Block Device, RBD

virtuelle Festplatte

direkt per Kernel eingebunden

- /dev/rbd0

als (Boot-) Image für KVM-VMs mit qemu+rbd

- libvirt
- Proxmox
- OpenStack
- u.a.

teilt Gigabyte-großes Image in viele kleine 4MB Ceph-Objekt auf

Gateway für iSCSI möglich

Rados-Gateway, S3, Swift

Object-Store mit HTTP-API

weitgehend kompatibel zu Amazon S3

Support für Object Locks

- WORM-ähnliches S3-Feature
- interessant für Archive / Backup

Nutzt mehrere Pools

- schneller SSD-Pool für Index
- platzsparender erasure coded HDD-Pool für Daten

CephFS

POSIX-kompatibles verteiltes Dateisystem

Client-Implementierung nur für Linux

- FUSE
- Kernel

Daten auf mehrere Pools aufteilbar

- schnelle oder langsame Pools für verschiedene Teile des Dateisystems

Quota, Snapshots

Dateisystem-Zugriffsrechte bestimmt der Client (!)

- ähnlich wie NFSv3 oder lokales Dateisystem

Gateways für SMB (Samba VFS) und NFS (Ganesha FSAL) möglich

FAQ

Kann ich mit Ceph meinen 300GB Fileserver ausfallsicher machen?

Nein.

Aber wenn Du Ceph brauchst, gibt es nichts anders.

Ceph 20 Tentacle

Ceph 20 Tentacle

Veröffentlicht am 18.11.2025

- aktuelle Version 20.2.1 vom 6.4.2026

Highlights:

FastEC

NVMe over TCP

SMB gateway

Case-insensitive directories in CephFS

Orchestrator

Seastore

Ceph 20 RADOS FastEC

Performance and space amplification optimizations (FastEC)

- Faktor 2 – 3

EC nun fast gleichauf mit Replikation => TCO reduziert

BlueStore mit besserer Kompression (lz4) und schnellerem WAL

https://ceph.io/en/news/blog/2025/tentacle-fastec-performance-updates/

Ceph 20 NVMe over TCP

NVMe-oF schneller als iSCSI

Orchestrator rollt NVMe-oF-Gateways aus

Konfiguration via CLI oder Dashboard

Multi-Subsystems, Multi-Namespaces und Multi-Pathing

Ein Namespace = Ein RBD-Image

Ceph 20 SMB

MGR module SMB

SMB single or cluster mode

Share configuration

https://ceph.io/en/news/blog/2025/smb-manager-module/

Verzeichniseinträge

Normalisierung

Groß/Kleinschreibung

https://docs.ceph.com/en/squid/cephfs/charmap/

Ceph 20 Orchestrator

mgmt-gateway

Ein Servicezugang für Dashboard und Monitoring-Stack

Unified Login via OAuth2

HA Service-IP möglich

certmgr

arbeitet als interne CA

überwacht Zertifikate

erneuert Zertifikate

Ceph 20 Crimson Seastore

Neuimplementierung des OSD

optimiert für NVMe

Tech-Preview verfügbar

https://ceph.io/en/news/blog/2025/crimson-seastore-vs-classic/

heinlein

Fragen und Diskussionen

Linux höchstpersönlich.

Bleiben wir im Kontakt

Robert Sander

Tel. +49 30 40 50 51-40
r.sander@heinlein-support.de

Heinlein Support GmbH
Schwedter Straße 8/9 | 10119 Berlin
www.heinlein-support.de