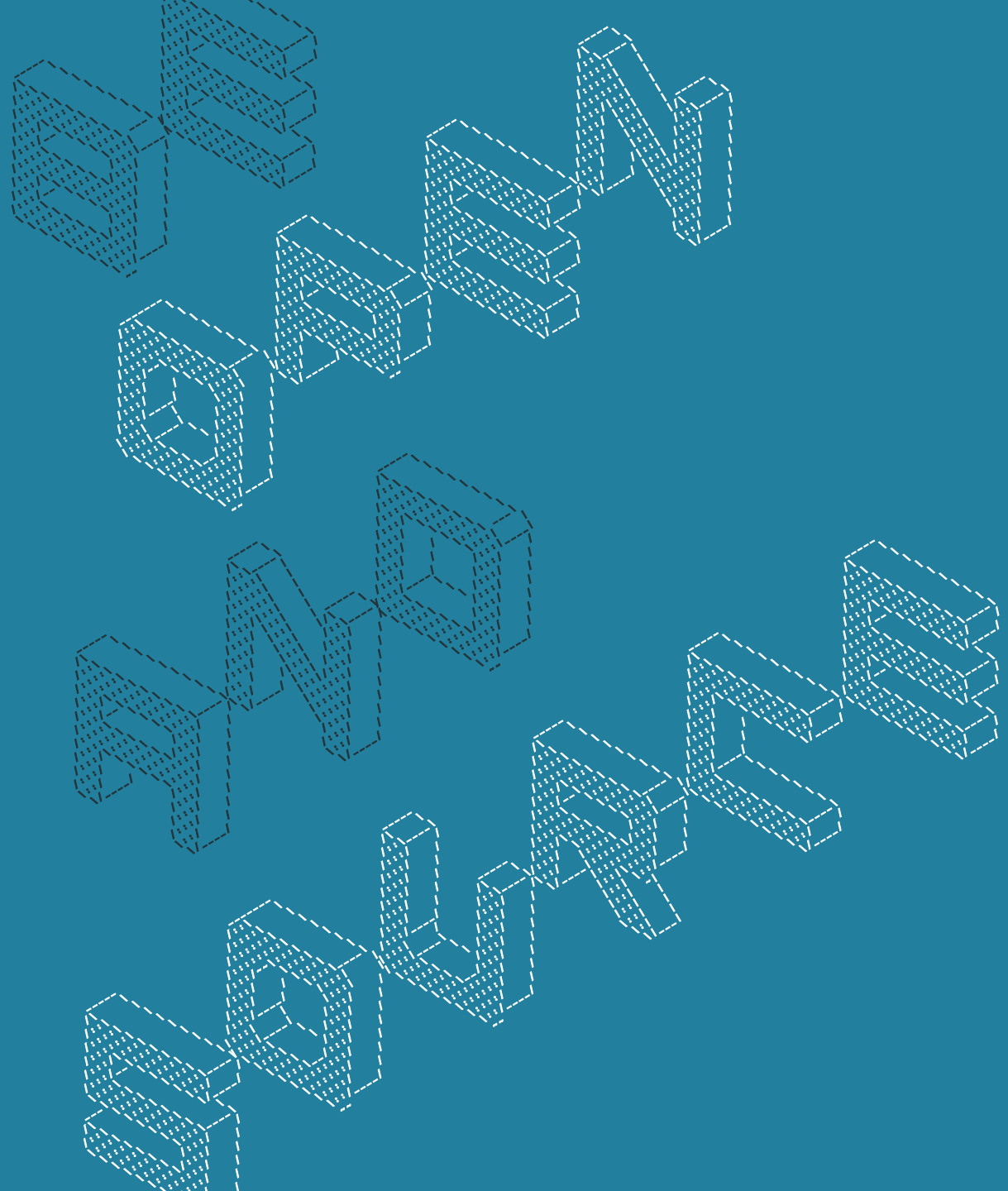


SLAC 2024

06.–08. Mai 2024 | Berlin

www.slac-2024.de

Einführung in Ceph

Software defined Storage

Abstraktion von Hardware

Hardware ist „egal“

Fing eigentlich schon mit LVM an

Beschränkt sich aber nicht nur auf eine Maschine

Redundanz nicht über RAID-Controller

Jede Hardware kann ausfallen

- Software natürlich auch

Skalierbarkeit

Beliebig in die Breite skalieren

Keine „teure“ vertikale Skalierung notwendig

günstigere Commodity Hardware einsetzbar

Trotzdem: Blick auf Performance wichtig

Ceph

Es war einmal eine Open Source Speicherlösung namens Ceph

Robert Sander | r.sander@heinlein-support | SLAC 2024 | 06.05.2024

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

Es war einmal eine Open Source Speicherlösung namens Ceph

Mit verschiedenen Komponenten, die zu kennen sich lohnt

Robert Sander | r.sander@heinlein-support | SLAC 2024 | 06.05.2024

Object Storage Daemon

OSD speichert Daten auf

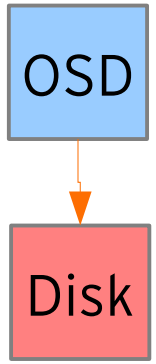
- HDD
- SSD
- NVMe
- oder was es noch geben wird

Ein Prozess pro Blockdevice

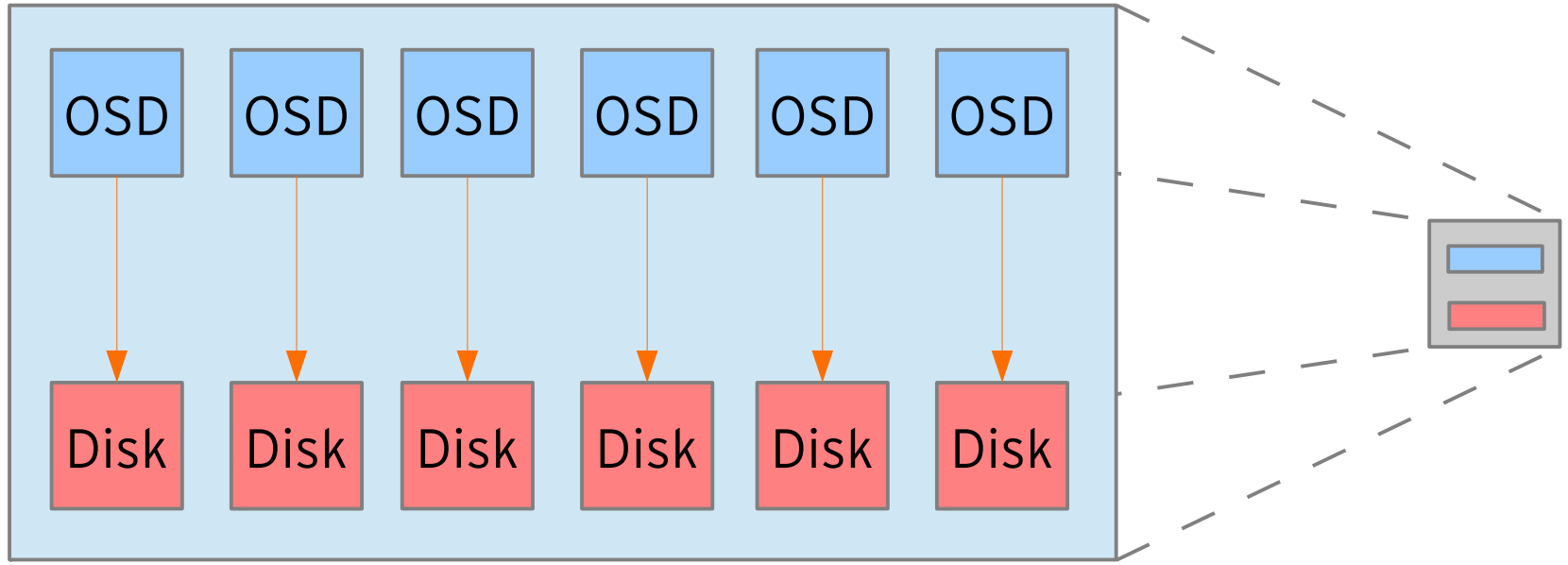
OSDs liefern Daten direkt an Clients

OSDs sprechen mit anderen OSDs

- Replikation
- Datenwiederherstellung



Mehrere OSDs in einem Ceph-Knoten



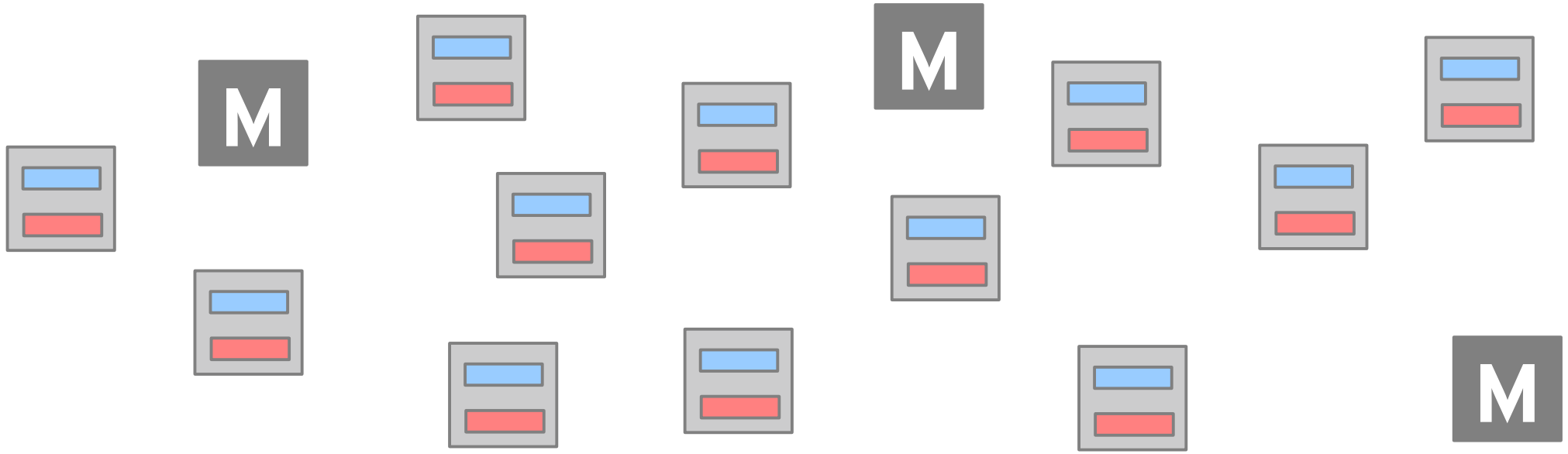
Dazu kommen Monitore

MONs bilden das Gehirn des Clusters

Quorum für Entscheidungen

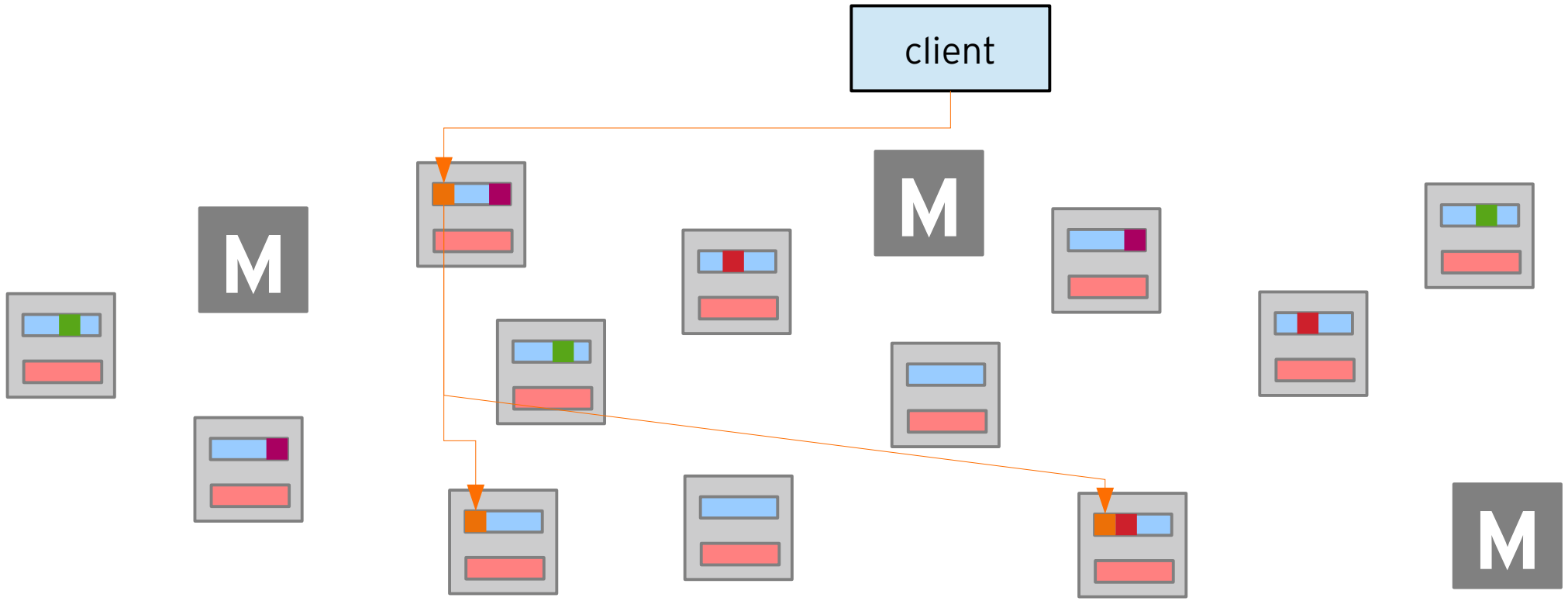
Nicht im Datenpfad

Und schon ist der Ceph-Cluster fertig



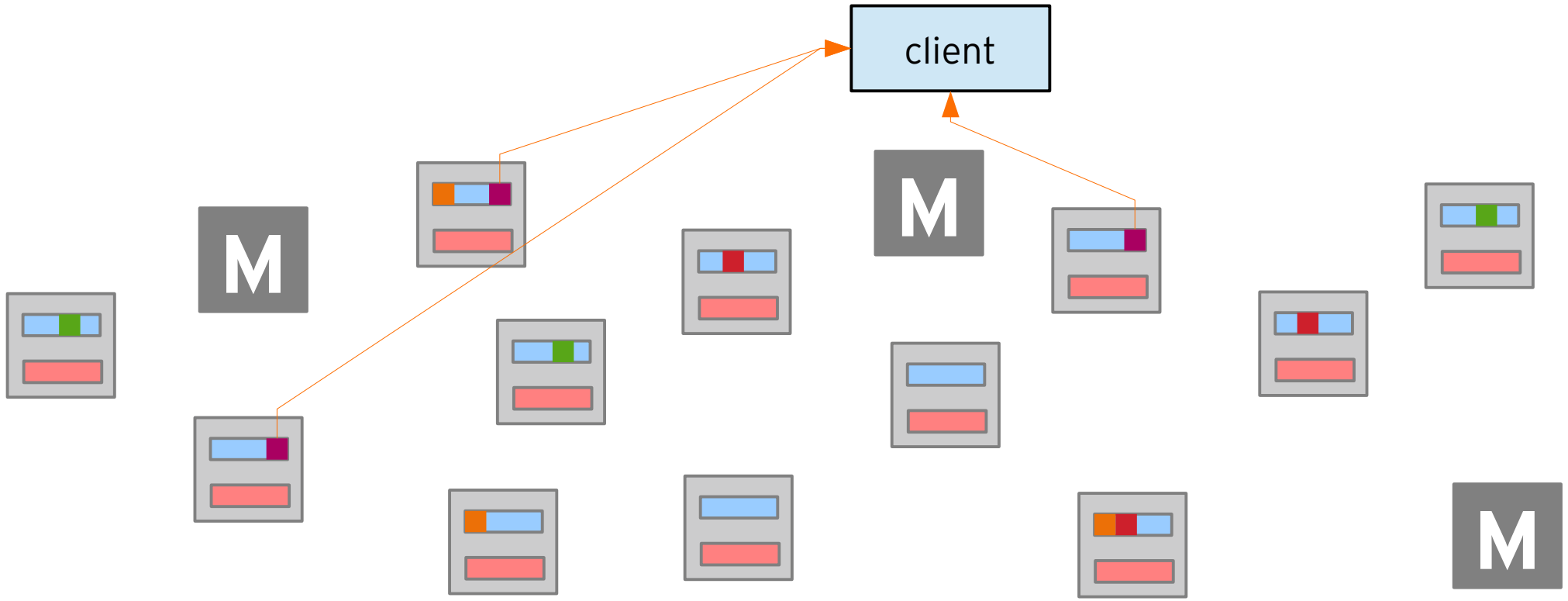
Schreiben …

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

Es war einmal eine Open Source Speicherlösung namens Ceph

Mit verschiedenen Komponenten, die zu kennen sich lohnt

Alle wollen Software Defined Storage anstelle von „Legacy“ Storage-Arrays

Robert Sander | r.sander@heinlein-support | SLAC 2024 | 06.05.2024

Klassische (proprietäre) Speichersysteme

Vorteile

- Einfach zu verstehen
- Vorhandene Erfahrung
- „Bauchgefühl“ besser
- Vorhersagbar in Verhalten und Kosten

Nachteile

- Streng kontrollierte Umgebung
- Begrenztes Wachstum
- Weniger Optionen
  - bestimmte HDDs / SSDs
  - Anzahl der Blockdevices
  - Netzwerk-Varianten
  - Controller
  - CPU

Software defined Storage

Vorteile

Selbermachen

Unbegrenztes Wachstum

Anpassungsfähigkeit

Auswahlmöglichkeiten

Nachteile

Selbermachen

Komplexität

Performance (Software CPU-bound)

Software defined Storage

Durchsatz

Latenz

IOPS

Verfügbarkeit

Zuverlässigkeit

Kapazität

Packungsdichte

Kosten

Konzipierung von SDS

Zielkonflikte

Verfügbarkeit gegen Packungsdichte

IOPS gegen Packungsdichte

Alles gegen Kosten

Große Auswahl an Hardware → Unübersichtlich

Softwareauswahl (gibt ja nicht nur Ceph)

Es gibt kein Standardrezept, das allen passt

Es war einmal eine Open Source Speicherlösung namens Ceph

Mit verschiedenen Komponenten, die zu kennen sich lohnt

Alle wollen Software Defined Storage anstelle von „Legacy“ Storage-Arrays

und hatten viele Fragen zu Konfigurationsmöglichkeiten

Robert Sander | r.sander@heinlein-support | SLAC 2024 | 06.05.2024

Netzwerk

die schnellste Netzwerktechnologie, die das Budget hergibt

Client-Zugriffs- und Cluster-Replikations-Netzwerk trennen

Replikations-Netzwerk mit mindestens Faktor 2 in der Bandbreite

Ethernet 10G < 40G < 25G < 100G

wegen der Latenz

Storage-Knoten

CPU, CPU, CPU

RAM, RAM, RAM

- 4GB pro OSD

guter Storage Controller

SSDs, SSDs, SSDs

HDDs

- günstiger
- eher für Archivdaten
- immer mit RocksDB+WAL auf SSD

Redundanz

Replikation

n genaue Kopien

hohe Leseraten

gute Schreibraten

- schnelles Cluster-Netzwerk

Wiederherstellung von mehreren Quellen

Netto-Kapazität nur 1/n

- n = 3 → 33%

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 + 2 unabhängige Knoten notwendig

Cluster ausbauen – Mehr Storage-Knoten

Gesamtkapazität steigt

Gesamtdurchsatz steigt

Gesamt-IOPS steigen

Verfügbarkeit erhöht sich

Latenz unverändert

Limit: Netzwerktopologie

Neuverteilung der Daten erzeugt temporär höhere Last

Es war einmal eine Open Source Speicherlösung namens Ceph

Mit verschiedenen Komponenten, die zu kennen sich lohnt

Alle wollen Software Defined Storage anstelle von „Legacy“ Storage-Arrays

und hatten viele Fragen zu Konfigurationsmöglichkeiten

und lernten, mit den richtigen Fragen Proof-of-Concepts zu bauen und auszubauen

Robert Sander | r.sander@heinlein-support | SLAC 2024 | 06.05.2024

Wie setze ich einen Ceph-Cluster zusammen?

Was soll der Cluster tun?

Budget abschätzen

Pilotprojekt bauen in ~10% der Zielgröße

Stellschrauben verändern, bis die Performance passt

Skalieren durch Hinzufügen weiterer Komponenten

Nicht von Anfang an perfekt, kann aber über die Zeit wachsen

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.

Es war einmal eine Open Source Speicherlösung namens Ceph

Mit verschiedenen Komponenten, die zu kennen sich lohnt

Alle wollen Software Defined Storage anstelle von „Legacy“ Storage-Arrays

und hatten viele Fragen zu Konfigurationsmöglichkeiten

und lernten, mit den richtigen Fragen Proof-of-Concepts zu bauen und auszubauen

und lebten glücklich bis ans Ende ihrer Tage

Robert Sander | r.sander@heinlein-support | SLAC 2024 | 06.05.2024

heinlein

Fragen und Diskussionen

Linux höchstpersönlich.

CompetenceCall

Das Backup für Ihre
Server-Administration.

Nutzen Sie unsere
SLA-Verträge und sichern
Sie sich den 24/7-Support unserer
Linux-Consultans.

- Kontimuierliche Absicherung mit garantierten Reaktionszeiten und festen SLAs
- Rückendeckung im Notfall: mindestens LPIC-2 zertifizierte Profis mit jahrelanger, täglicher' Admin-Erfahrung
- Projektunterstützung: maßgeschneiderte Lösungen, die Flexibilität, Sicherheit, Administrierbarkeit und Hochverfügbarkeit vereinen
- Services: Performanceanalyse, Serverhärtung, Netzwerkanalyse, Konfigurationshilfe, Datenrestaurierung

Werde Teil des Teams

- Du bist neugierig, voller Tatendrang und überzeugt von Linux, Open Source und sicherer, freier Kommunikation?

- Wir freuen uns über Unterstützung im Team:
  www.heinlein-support.de/jobs

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