Software Defined Storage

15INM
Michael Horn & Andrej Lisnitzki

Cloud Computing

- internetzentriertes Entwicklungsansatz
- Bereitstellung komplexen Leistungen aus Soft- und Hardware in Form eines Abstrakten Dienstes
  - Speicher, Rechenzeit
  - Anwendungssoftware als Service über das Internet
  - Bereitstellung der komplexeren Dienste über festgelegte Schnittstellen
- Unabhängigkeit von der Hardware, auf der die Leistungen laufen

Cloud Computing

(technisches Cloud-Stack)

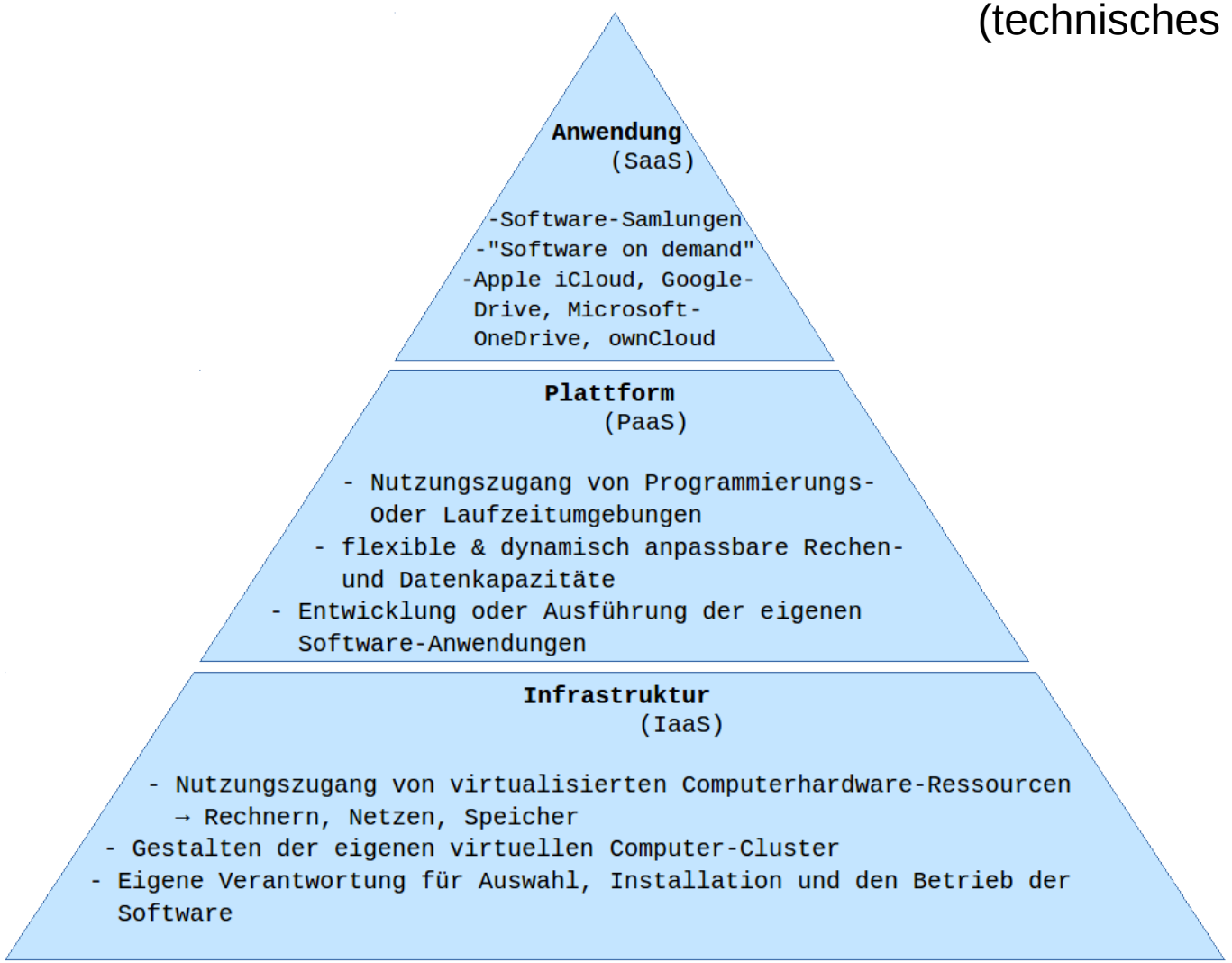


Abbildung: technische Realisierung von Cloud Computing

Cloud Computing

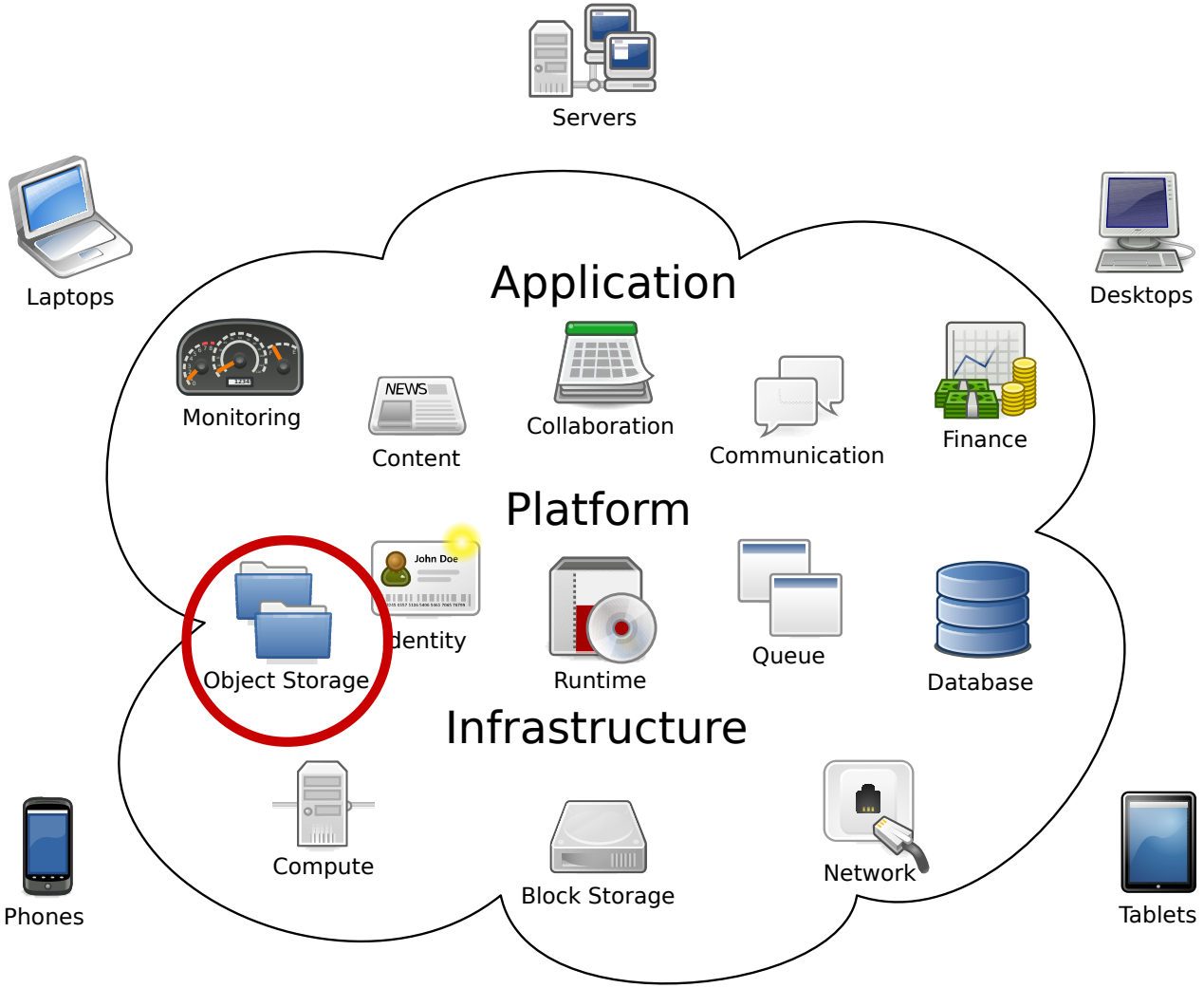


Abbildung: Elemente des Cloud Computing [https://de.wikipedia.org/wiki/Cloud_Computing (2017.01.16)]

Skalierbarkeit

Ist die Fähigkeit eines Systems die Leistung durch das Hinzufügen der Ressourcen zu steigern.

Abbildung: Skalierbarkeit
[http://www.staffsupply.at/images/fotolia/Technische_Daten_Skalierbarkeit.jpg (2017.01.16)]

Skalierung

- Vertikale Skalierung (scale up)
  - Steigerung der Leistung durch das Hinzufügen der Ressourcen zu einem Knoten des Systems

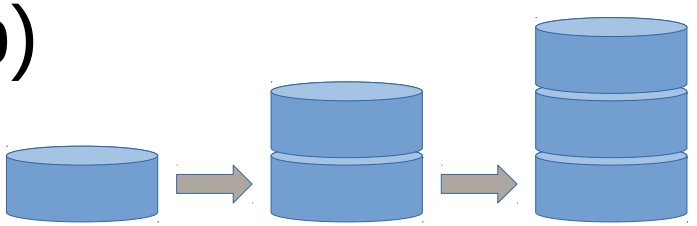


Abbildung: Vertikale Skalierung

- Horizontale Skalierung (scale out)
  - Steigerung der Leistung eines Systems durch das Hinzufügen zusätzlicher Knoten

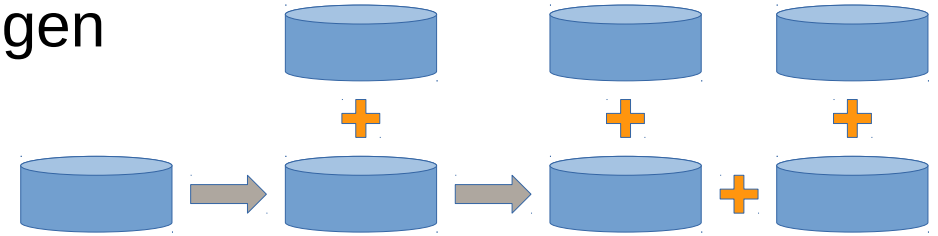


Abbildung: Horizontale Skalierung

Blockgeräte

- „speichern“ Blöcke fester Größe
  - Blöcke besitzen eine eindeutige Adresse
  - Blockgrößen sind typischerweise Zweierpotenzen
  - Jeder Block kann individuell für ein r/w -Zugriff adressiert werden
- Festplatten, Disketten, CD-ROMs

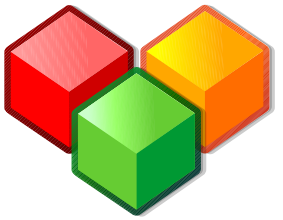


Abbildung: Linux-Zeichen für ein Blockgerät
[https://de.wikipedia.org/wiki/Ger%C3%A4tedatei (2017.01.16)]

Datei-, Block-, Objektspeicher

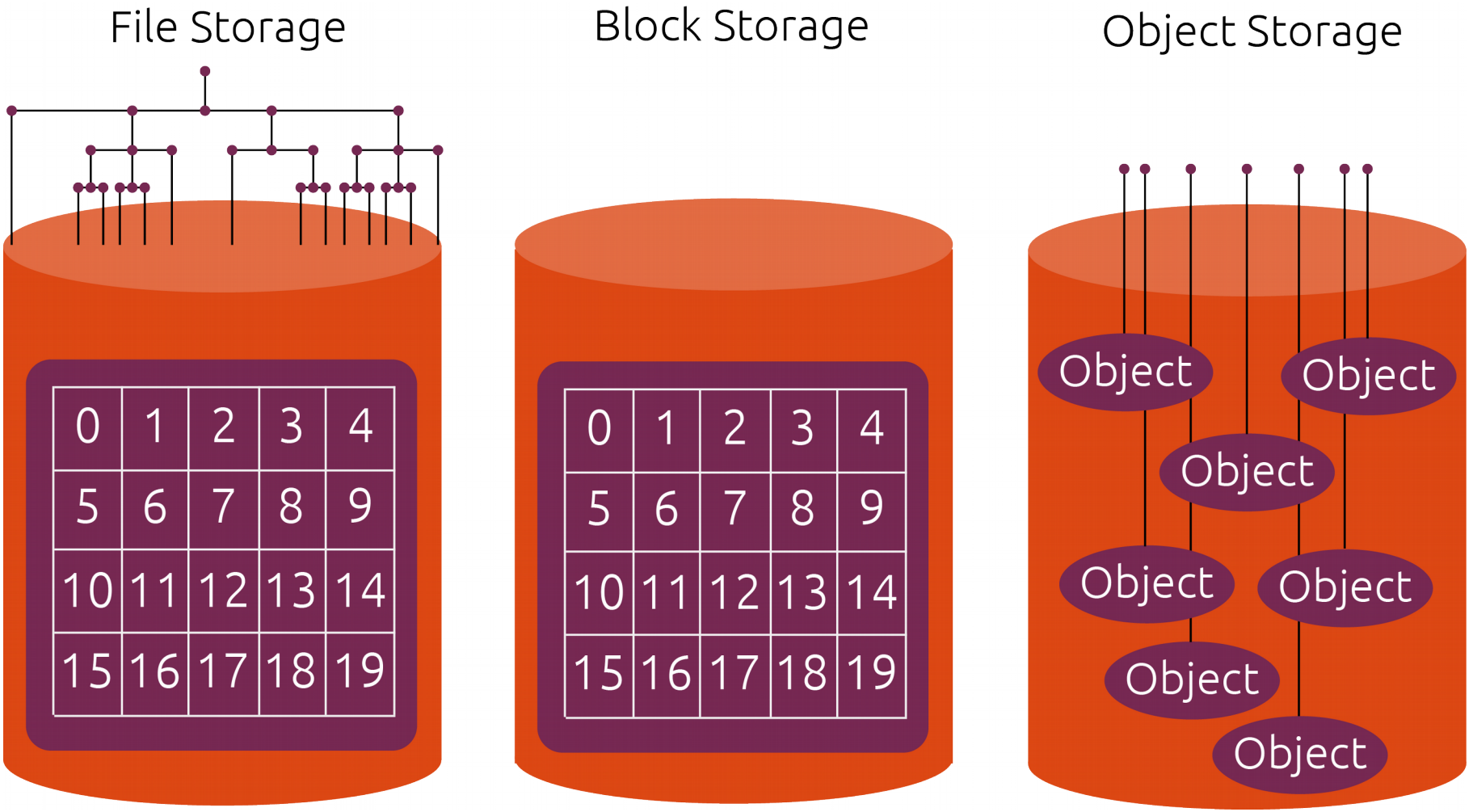


Abbildung: Datei-, Block-, Objektspeicher
[https://insights.ubuntu.com/2015/05/18/what-are-the-different-types-of-storage-block-object-and-file/ (2017.01.16)]

Dateispeicher

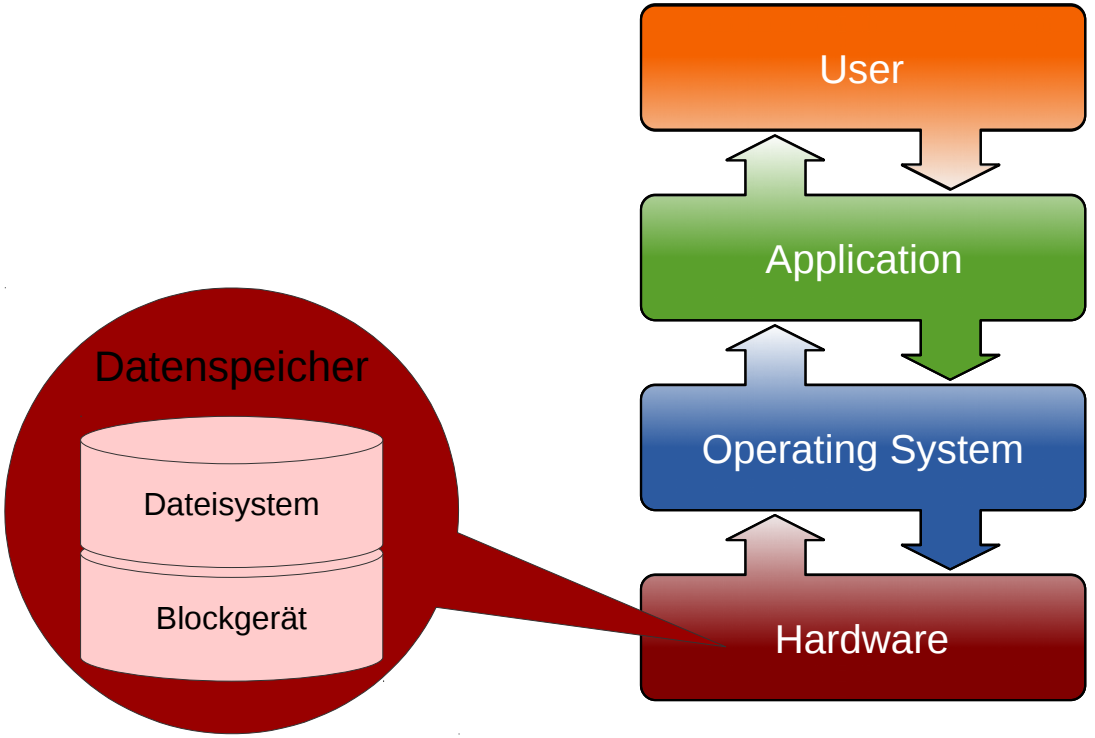


Abbildung: Schematische Zuordnung von Datenspeicher in einem System

Objektspeicher

- Dateien als Objekte

  Enthalten:

  - Daten
  - globale eindeutige Kennung
  - Metainformation → zur besseren Auffindbarkeit
- Für unstrukturierte Daten geeignet

  Medien, Dokumente, Protokolle, Backups,

  Anwendungsbinärdaten, VM-Images
- Deduplikation und Replikation
- Einfache Erweiterbarkeit
- Betriebssystem-Neutralität

→ Facebook, Spotify, Dropbox etc.

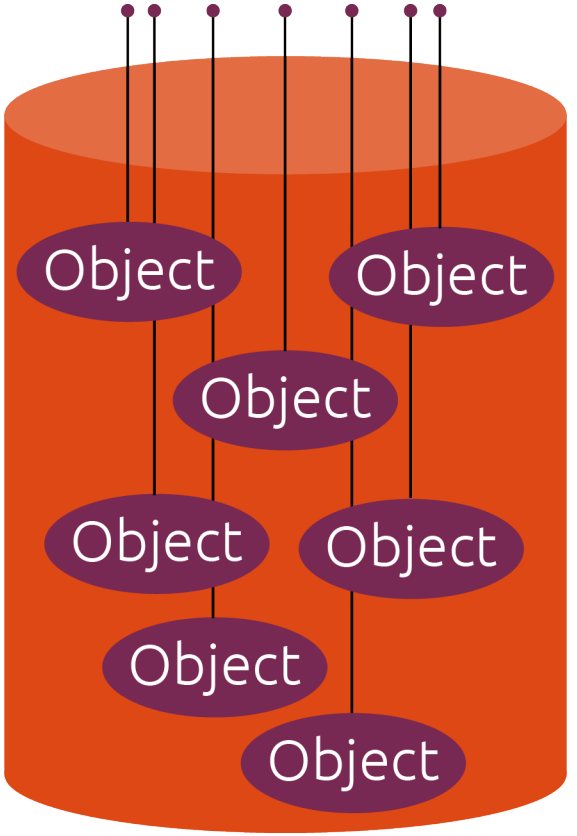


Abbildung: Schematische Darstellung vom Objektspeicher

Speicher-Vorrichtungen - DAS

DAS (Direct Attached Storage)

- An einzelnem Rechner angeschlossene Festplatten
- SCSI, SAS & andere blockorientierte Übertragungsprotokolle

Vorteile:

- Geringer Hardwareaufwand
- Sehr hohe Datenübertragungsrate, da direkt an das System Angeschlossen (von der verwendeten Technik abhängig)
- Kein zusätzlicher Protokollstack

Nachteile:

- Nur an ein Host gebunden → begrenzt in der Skalierung
- Andere Computer können auf die DAS-Festplatten nur über den Rechner, an dem diese physisch angeschlossen sind, per Netzwerk nutzen.

Speicher-Vorrichtungen - SAN

SAN (Storage Area Network)

- Ein Netzwerk zur Anbindung von Festplattensubsystemen an Serversysteme
- SCSI-Kommunikationsprotokoll per Fibre-Channel o. iSCSI (Transport-Protokoll)

Vorteile:

- Hohe Datenübertragungsrate (dank Fibre-Channel)
- Speicher-System kann von dem Server entfernt werden (Glasfaser: bis zu 30km)

Nachteile:

- Hardware der unterschiedlichen Hersteller ist nicht immer untereinander kompatibel

Speicher-Vorrichtungen - NAS

- **NAS** (Network Attached Storage)

  Ist ein Serverdienst, die den über einen Netzwerkdienst angeschlossenen Clients betriebssystemabhängig einsatzbereites Dateisystem zur Verfügung stellt.

  Vorteile:

  - Geringere Energieverbrauch
  - Für Endnutzer komfortable Nutzung durch schon bestehende Dateibasierte-Dienste (z.B. NFS)
  - Zusätzliche Funktionalitäten möglich (Druckserver, Mail-Benachrichtigung, etc.)

  Nachteile:

  - Relativ langsame Datenübertragungsrate
  - Begrenzt in der Skalierung

Speicher-Vorrichtungen

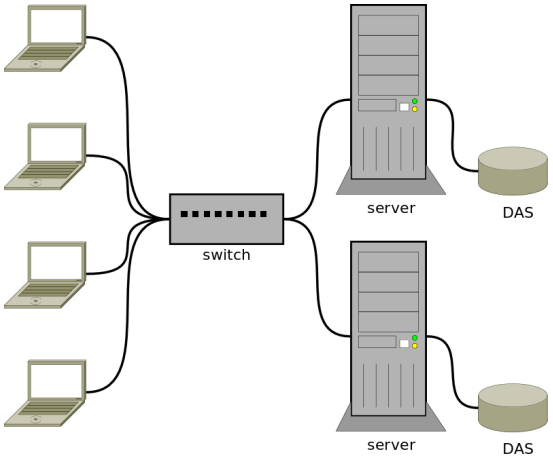


Abbildung: DAS

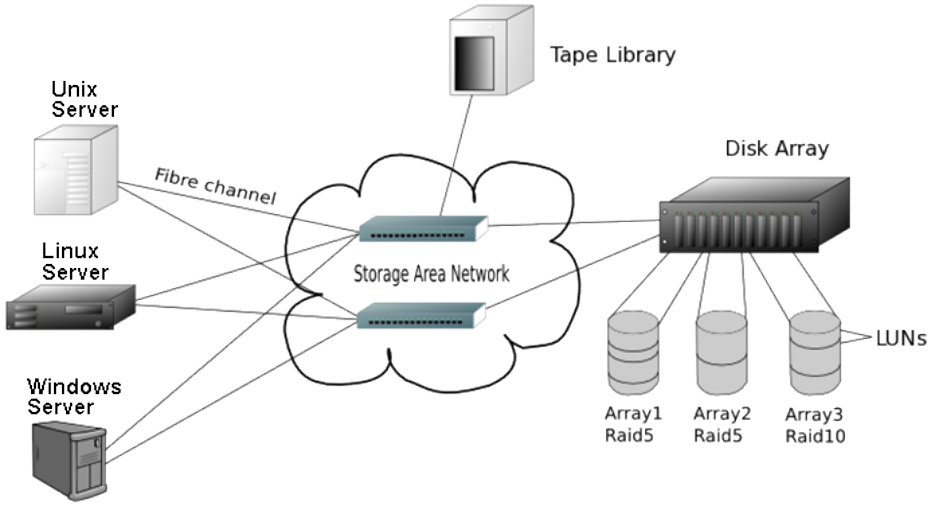


Abbildung: SAN [https://de.wikipedia.org/wiki/Storage_Area_Network (2017.01.16)]

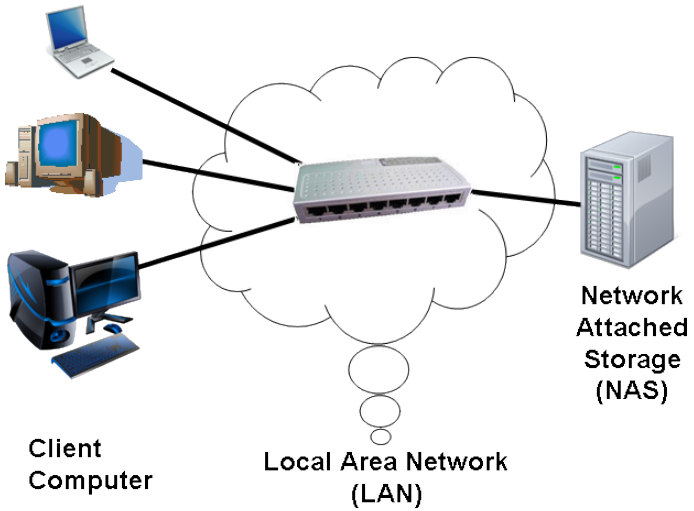


Abbildung: NAS [https://de.wikipedia.org/wiki/Network_Attached_Storage (2017.01.16)]

Speicher-Vorrichtungen

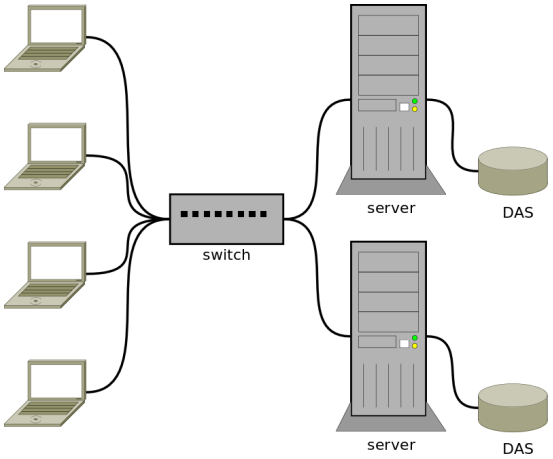


Abbildung: DAS

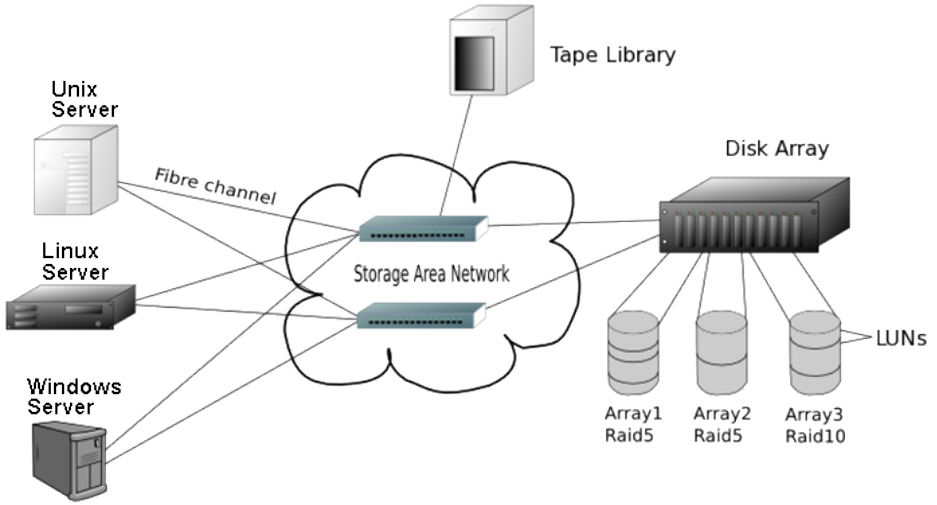
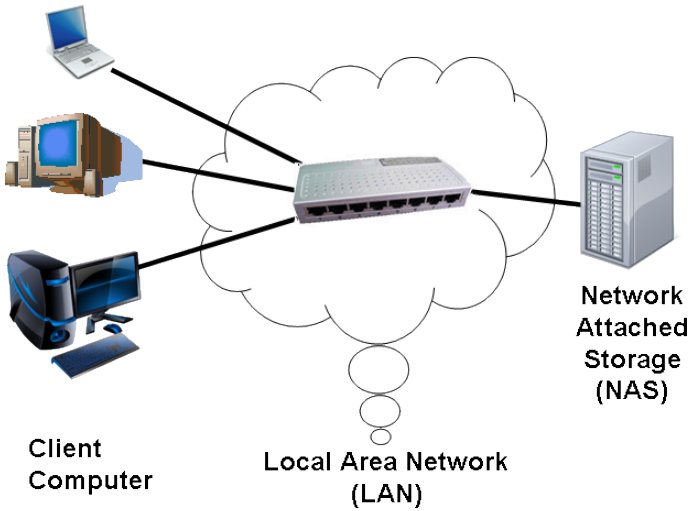


Abbildung: SAN [https://de.wikipedia.org/wiki/Storage_Area_Network (2017.01.16)]



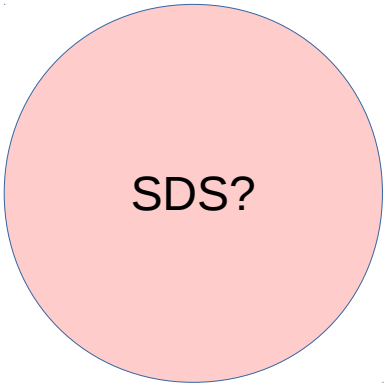


Abbildung: NAS [https://de.wikipedia.org/wiki/Network_Attached_Storage (2017.01.16)]

Was ist Software Definied Storage?

- Objektorientierter Speicheransatz
- Jeder Speicherknoten ist für Speicherung einer Teilmenge der Gesamtdaten verantwortlich
- SDS-Lösungen fügen eine zusätzliche Ebene zwischen physischen Datenträgern und Frontend ein
- Zusätzliche Ebene verteilt im Hintergrund Daten (bei Schreibzugriffen)

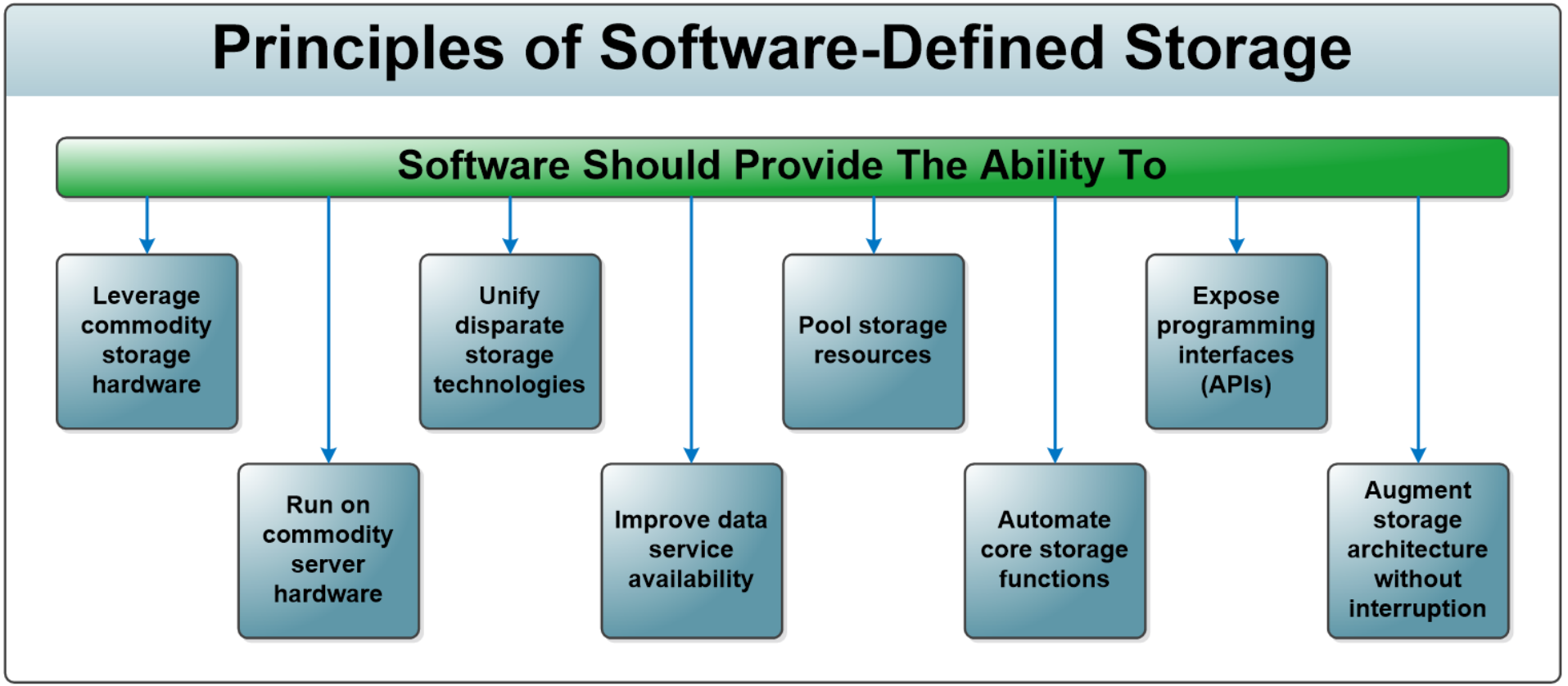
Was ist SDS?

- SDS-Lösungen betrachten Daten als zerlegbare binäre Objekte
- Beim Auslesen werden zerlegte Objekte wiederhergestellt
- Namensgebend für SDS ist diese Ebene, denn:
  - Eigentliche Speicherlösung ist als Software implementiert

Was ist SDS?

- Zugriff ist über Frontend möglich
- Frontends können als unterschiedliche Typen realisiert sein
- Frontend könnte ein Blockgerät nachbilden
  - System sieht es als Blockgerät und kann damit normal arbeiten, aber im Hintergrund läuft SDS
- Vielzahl von Frontends möglich

Prinzipien von SDS



Vorteile/Nachteile

- Sehr flexibel in der Skalierung
  - Größere Speicherkapazität
  - Höhere Verfügbarkeit
  - Erhöhten Datendurchsatz
- Leistungsfähige Netzwerkinfrastruktur benötigt

Ein Überblick am Beispiel:

Beispiel Video

Ceph Promo

Wer verwendet Ceph?

Geschichte

- Erfinder Sage Weil (geboren 17.03.1978)
- Entstand im Rahmen einer Doktorarbeit (~2006) an der University of Santa Cruz in Kalifornien
- Aufgabe war es für das US Department of Energy eine Storage-Lösung zu entwickeln, welche die Nachteile klassischer Systeme nicht hat
- Am Anfang der Entwicklung stand das Ziel, ein mit POSIX kompatibles Dateisystem zu erstellen
- Gründete später das Unternehmen Inktank Storage (aufgekauft von Red Hat)

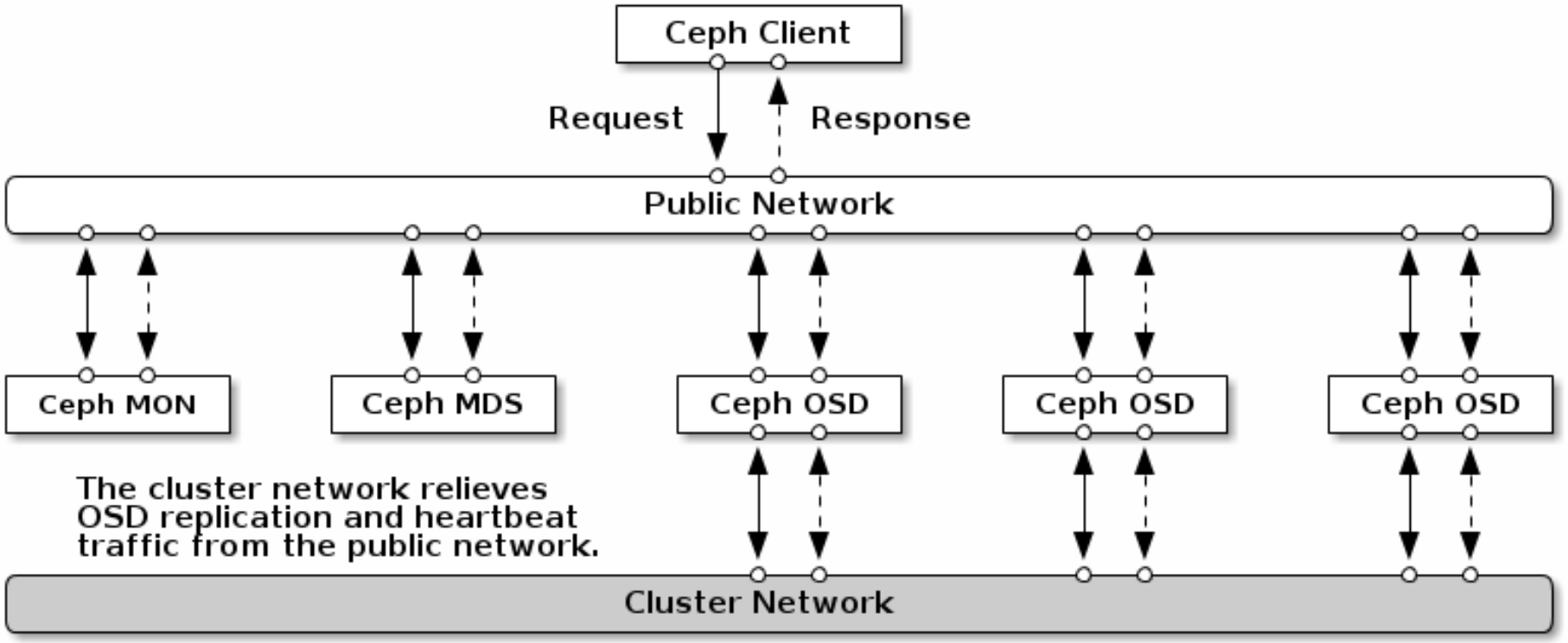
Was ist Ceph?

- Ceph ist ein klassischer Objektspeicher für SDS
- Anwendung hieß ursprünglich RADOS
- RADOS-Kürzel beschreibt Funktionsweise von SDS/CEPH:
  - Reliable Autonomic Distributed Object Store
    (Zuverlässige Automatische Verteilte Objekt Speicherung)
- Wegen Cloud-Boom umbenannt in Ceph
- “Ceph” als Abkürzung für Cephalopoda (Kopffüßler)
  - Synonym für die Arbeitsweise des SDS

Aufbau

- Typisches Ceph-Cluster besteht aus zwei Diensten, die innerhalb eines Clusters beinahe beliebig oft vorkommen:
  - Object Storage Devices (OSD)
  - Monitoring Server (MON)
- In CephFS (Frontend) existiert zusätzlich noch:
  - Meta Data Storage (MDS)

Aufbau


Ceph Client
Request
Response
Public Network
Ceph MON
Ceph MDS
Ceph OSD
Ceph OSD
Ceph OSD
The cluster network relieves
OSD replication and heartbeat
traffic from the public network.
Cluster Network

OSD (Object Storage Device)

- Dienst, der physische Platten in Clusterverbund integriert
- Ebene wodurch SDS definiert wird stellen genau diese OSDs dar
- Kommunikation mit den physischen Blockgeräten wickeln die OSD-Server im Hintergrund ab
- Auch um die inhärente Replikation kümmern sich die OSD-Dienste

OSD (Object Storage Device)

- Pro physischem Blockgerät läuft ein OSD-Dienst
  - Zum Beipiel: Auf Server mit 10 Festplatten laufen 10 OSD-Dienste
- OSDs bilden Anlaufpunkt für Clients
- Clients liefern Daten bei einem OSD ab, das sich dann um Replikation kümmert
- Erst wenn Replikationsvorgaben erfüllt sind, erhält der Client die Nachricht, dass der Schreibvorgang erfolgreich war

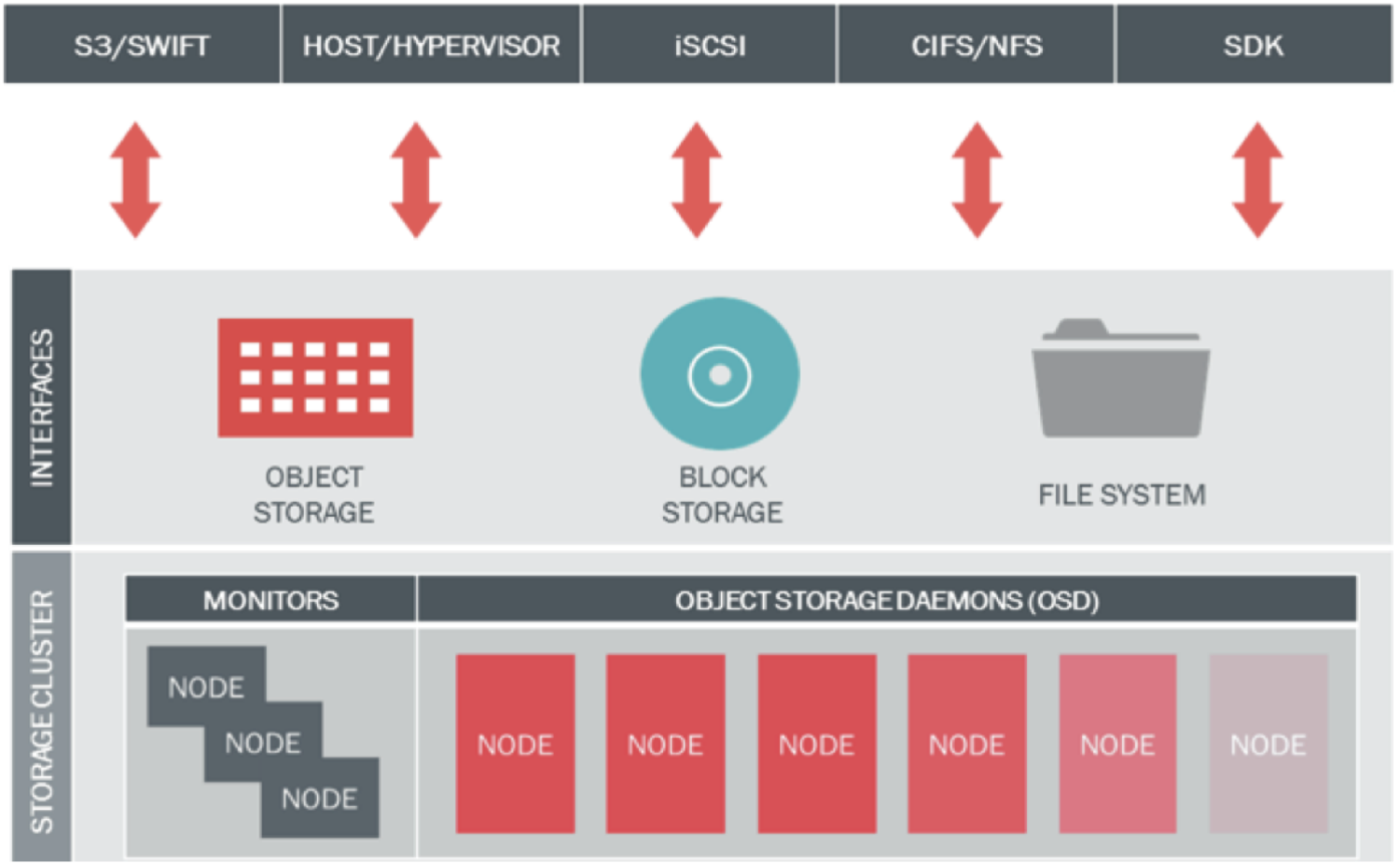
MON (Monitoring Server)

- Sind die Cluster-Wachhunde
- Ein MON überwacht mehrere OSDs
- Sie führen Buch über vorhandene MONs und OSDs und erzwingen im Cluster ein Quorum [Verfahren zur Gewährleistung der Datenintegrität] auf der MON-Ebene
- Wichtig in Situationen, in denen das Cluster in mehrere Partitionen zerfällt, etwa weil Netzwerkhardware kaputtgeht

MON (Monitoring Server)

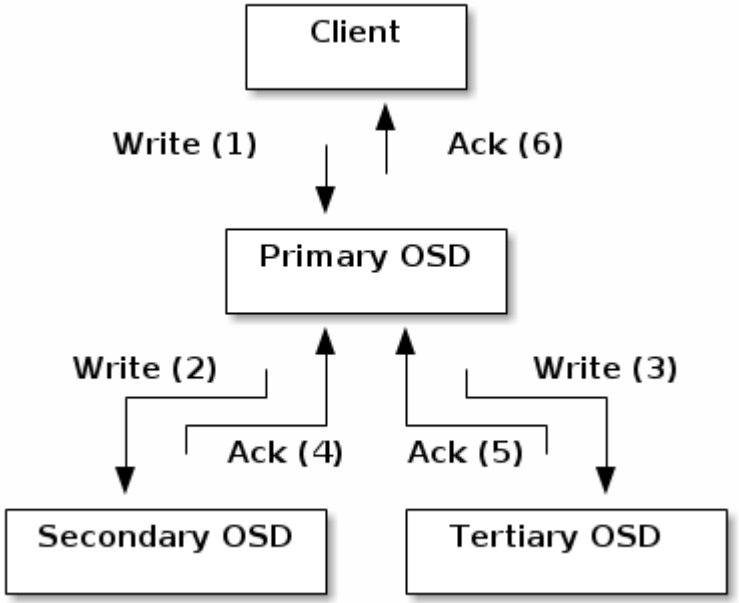
- MONs stellen in solchen Szenarien sicher, dass Clients nur auf Partition des Clusters schreibend zugreifen können, die die Mehrheit der insgesamt im Cluster bekannten MON-Server hinter sich weiß
- Abfragen der Dienste erfolgt durch regelmäßige Hearbeats
- MON-Server ist Anlaufpunkt für Clients/OSDs, wenn Informationen über aktuelle Topologie des Clusters benötigt werden

Aufbau


S3/SWIFT
HOST/HYPERVISOR
iSCSI
CIFS/NFS
SDK
INTERFACES
OBJECT STORAGE
BLOCK STORAGE
FILE SYSTEM
STORAGE CLUSTER
MONITORS
OBJECT STORAGE DAEMONS (OSD)
NODE
NODE
NODE
NODE
NODE
NODE
NODE
NODE
NODE

Datenverteilung

- OSDs tragen Verantwortung, dass eingehende Daten auf physischen Speichergeräten sinnvoll verteilt und abgelegt werden (Write)
- Read erfolgt in umgekehrter Reihenfolge



Datenverteilung

- Woher aber erfährt ein Client, der Daten in das Cluster laden möchte, welches OSD das richtige ist?
- Wo bekommt das angesprochende OSD die Information her, auf welchen anderen OSDs es von den hochgeladenen Daten Replikate anlegen muss, bevor es dem Client einen erfolgreichen Schreibvorgang vermeldet?
  - CRUSH-Algorithmus
  - Cephs Verteilungs-Algorithmus und ermöglicht es Clients wie OSDs, sich das jeweils passende Ziel-OSD für bestimmte Datensätze auszurechnen

CRUSH-Algorithmus

- Algorithmus zum Platzieren von Daten
- Steht für Controlled Replication Under Scalable Hashing
- Gemeint ist Prinzip, anhand dessen Clients oder OSDs die Ziel-OSDs für bestimmte Objekte festlegen
- Wenn ein Client Daten in das Cluster laden möchte, findet zuerst die Aufteilung in Objekte statt

CRUSH-Algorithmus

- Für jedes der Objekte stellt sich dann die Frage, an welches OSD es zu senden ist und wohin es von dort repliziert wird
- Client organisiert von MON-Servern des Ceph-Clusters das aktuelle Verzeichnis aller OSDs und stößt dann die Crush-Berechnung für das jeweilige Objekt an
- CRUSH lässt sich von außen beeinflussen

CRUSH-Algorithmus

- CRUSH-Map bietet Admin die Möglichkeit, Rechner oder OSDs logisch zu gruppieren
  - Legt Admin zum Beispiel fest, dass bestimmte Server in Rack 1 hängen und andere Server in Rack 2, so kann er bestimmen, dass Replikate eines Objektes in beiden Racks vorhanden sein müssen
- Wenn Client oder OSDs die Crush-Kalkulation für ein bestimmtes Objekt durchführen, beziehen sie die CRUSH-MAP mit ein
- Solange sich Topologie des Clusters nicht ändert, bleibt das CRUSH-Resultat identisch

Parallelität

- Größte Stärken von Ceph ist, dass CRUSH-Kalkulationen ,Objekt-Uploads oder Objekt-Downloads parallel geschehen
- Client zerteilt eine Datei in viele kleine Objekte und lädt diese parallel auf verschiedene OSDs
- Client redet immer mit vielen gleichzeitig und kombiniert so die Bandbreite beim Hoch- oder Herunterladen
- Verglichen mit klassischen Storages erreicht Ceph bei entsprechender Netzwerkhardware enorme Durchsatzwerte

Weiterführende Literatur

- Ceph ist Open Source
  - https://github.com/ceph
- Wissenschaftliche Arbeiten von Sage Weil zu finden unter:
  - http://ceph.com/resources/

Zusammenfassung

- Ceph ist ein in die horizontale skalierendes SDS
- Abstrahiert Zugriffe auf physischen Speicher durch zusätzliche Ebene
- Zerlegt Daten in mehrere Binärobjekte und verteilt sie auf verschiedene Blockgeräte
- Kann hohe Performance durch Parallelität erreichen
- Zugriff per Frontends

Quellen

- https://en.wikipedia.org/wiki/Sage_Weil
- https://de.wikipedia.org/wiki/Ceph
- http://www.golem.de/news/cloud-computing-was-ist-eigentlich-software-defined-storage-1610-122478.html
- http://docs.ceph.com/docs/master/rados/
- http://www.bitoss.com/valueaddedservices/ceph/
- http://www.searchstorage.de/definition/Objekt-Storage-Object-Storage
- https://insights.ubuntu.com/2015/05/18/what-are-the-different-types-of-storage-block-object-and-file/
- http://os.inf.tu-dresden.de/~ch12/sub/diplom/node14.html
- http://www.golem.de/news/cloud-computing-was-ist-eigentlich-software-defined-storage-1610-122478-2.html
- https://ru.wikipedia.org/wiki/Software-defined_storage
- https://de.wikipedia.org/wiki/Direct_Attached_Storage
- https://de.wikipedia.org/wiki/Storage_Area_Network
- https://de.wikipedia.org/wiki/Network_Attached_Storage
- https://de.wikipedia.org/wiki/Cloud_Computing
- https://de.wikipedia.org/wiki/Skalierbarkeit
- https://en.wikipedia.org/wiki/Device_file