Lustre

Proseminar „Speicher- und Dateisysteme”

Marten Backmann

Arbeitsbereich Wissenschaftliches Rechnen
Fachbereich Informatik
Fakultät für Mathematik, Informatik und Naturwissenschaften
Universität Hamburg

5. Dezember 2017

Gliederung (Agenda)

Einleitung

Was ist Lustre?

- Name leitet sich von Linux und Cluster ab
- paralleles, verteiltes Dateisystem
- hoch skalierbar
- verfügbar unter GNU GPL v2

Geschichte [5]

- Entstand aus einem Projekt an der Carnegie Mellon University
- 2001: Cluster File Systems, Inc
- Entwicklung von Lustre im Auftrag des Energieministeriums (DEO)
- 2007: Sun Microsystems kauft Cluster File Systems
- 2010: Oracle kauft Sun Microsystems
- 2013: Xyratex Ltd. kauft die Rechte von Lustre von Oracle
- 2015: lustre.org

Release History [5, 2]

- Version 2.0, August 2010
  - Vorbereitung auf große Veränderungen der Architektur
- Version 2.1, September 2011
  - Antwort auf das Einstellen der Entwicklung bei Oracle
  - Vergrößerung der maximalen OST-Größe von 24 auf 128 TB
  - Abwärtskompatibel zu Version 1.8.6 Clients
- Version 2.2, März 2012
  - bessere Metadata-Performance, gleichzeitiges Arbeiten in einem Verzeichnis
  - Eine Datei kann auf bis zu 2000 OSTs verteilt werden
- Version 2.3, Oktober 2012
  - Vorläufige Unterstützung für ZFS

- Version 2.4, Mai 2013
  - Distributed Namespace (DNE)
  - ZFS für MDTs und OSTs
- Version 2.5, Oktober 2013
  - Hierarchical Storage Management (HSM)
- Version 2.6, Juli 2014
  - Vorläufige Unterstützung verteilter Verzeichnisse
- Version 2.7, März 2015
  - verbesserte Unterstützung verteilter Verzeichnisse
- Version 2.8, März 2016
  - Fertige Unterstützung verteilter Verzeichnisse
- Version 2.9, Dezember 2016
  - Mounten von Unterverzeichnissen
  - Verschlüsselung der Client-Server-Kommunikation
- Version 2.10, Juli 2017
  - LNet Multi-Rail
  - Snapshots

Skalierbarkeit [3, 4]

	theoretisch möglich	in der Praxis
Clients	100-100.000	50.000+
Dateigröße	32 PB oder 2^{63} bytes $\sim$ 9,22 EB	mehrere TB
Systemgröße	512 PB	55 PB
OSS-Skalierbarkeit	1-32 OSTs pro OSS 1000 OSSs mit bis zu 4000 OSTs	450 OSSs mit 1000 4TB OSTs
OSS-Performance	15 GB/s	10 GB/s
MDS-Skalierbarkeit	max 256 MDS mit je 256 MDTs 4 Mrd Dateien (ldiskfs) 64 Mrd Dateien (ZFS)	3 Mrd Dateien
MDS-Performance	200.000/s stat-Operationen	50.000/s

Architektur

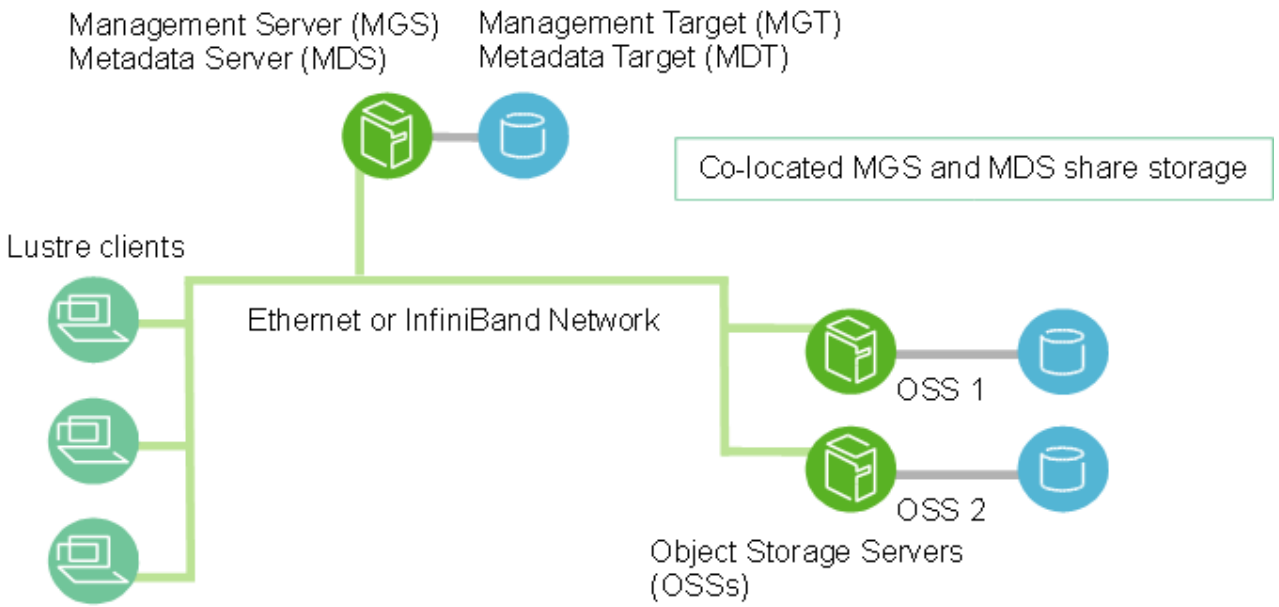

Abbildung: (fast) minimaler Aufbau [2]

Management Server (MGS) und Target (MGT)

- „Globales Register"
- Verwaltet Konfigurationen für Clients, Server und Targets
- Nimmt nicht direkt an den Operationen im Dateisystem teil

Metadata Server (MDS) und Target (MDT)

- Speichert Dateinamen, Pfad Zugriffsrechte und Datei-Layout
- Metadaten liegen auf MDT
- Verlust des MDT kritisch
- Bis zu 256 MDTs (pro MDS) möglich

Object Storage Server (OSS) und Target (OST)

- OSS: I/O-Service für einen oder mehrere OSTs
- Dateien bestehen aus einem bis mehreren Objekten
- Objekte sind über alle OSTs hinweg verteilt

Object Storage Device (OSD)

- Lustre-Targets besitzen ein eigenes Dateisystem:
  - ldiskfs
  - ZFS (Zeta File System)
- Targets können ünterschiedliche Dateisysteme besitzen

Lustre-Client

- Kombiniert Metadaten- und Objektspeicher zu einem kohärenten POSIX-Dateisystem
- Anwendungen müssen daher nicht angepasst werden
- Management Client (MGC)
- Metadata Client (MDC)
- Object Storage Client (OSC)

Lustre-inodes

- beinhaltet sämtliche Metadaten einer Datei
- von ext4 abgewandelt
- Extended Attributes (EA)
- FID

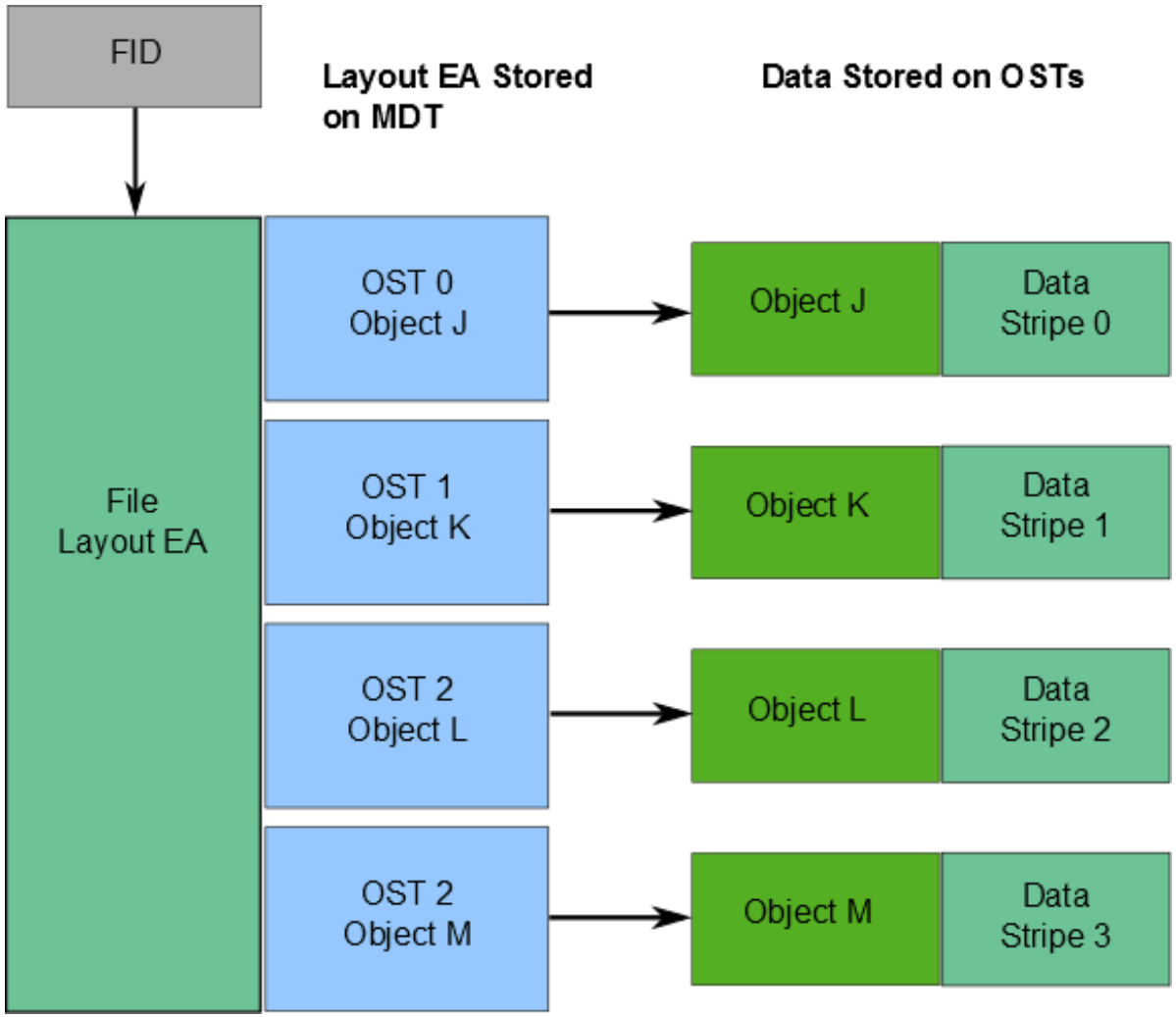

Abbildung: layout EA [4]

Striping

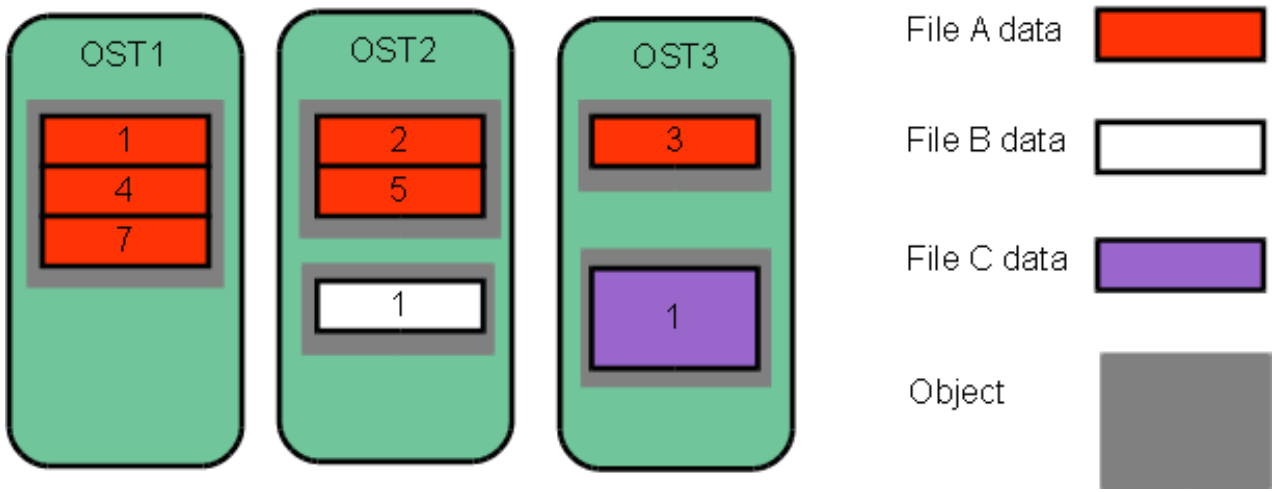

Abbildung: Striping [1]

- Hoher Durchsatz durch das Verteilen einer Datei auf mehrere Targets
- Höherer Overhead
- erhöhte Wahrscheinlichkeit von Datenverlust

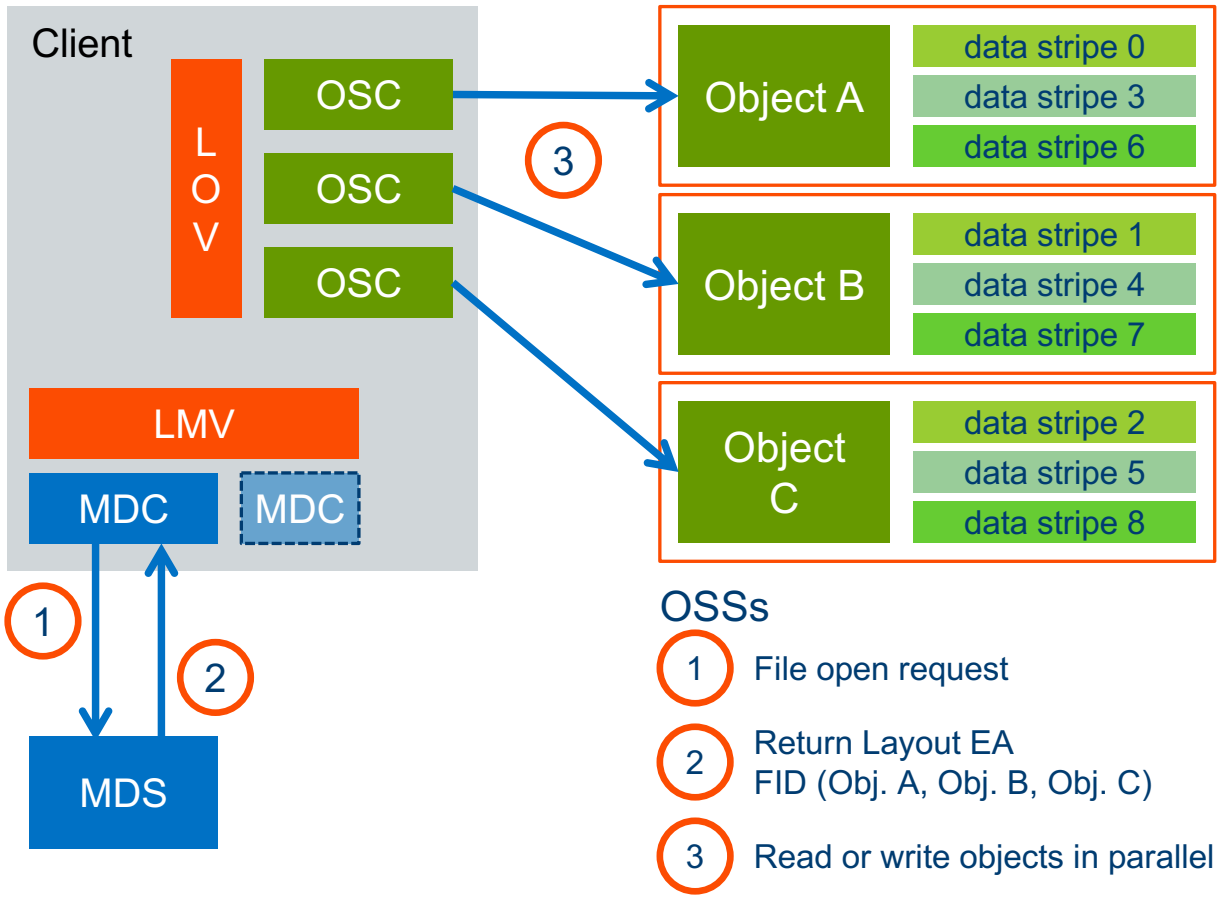

Abbildung: Ablauf [3]

Zusammenfassung

- Aufteilung von Daten und Metadaten
  - Metadata Server u. Target
  - Object Storage Server u. Target
- Hohe Skalierbarkeit
- Striping
- POSIX-konform

Quellen

[1] Oak Ridge National Laboratory. Lustre® basics. URL: https://www.olcf.ornl.gov/kb_articles/lustre-basics/ [cited 2017-12-4].

[2] OpenSFS. Lustre wiki. URL: http://wiki.lustre.org [cited 2017-12-4].

[3] OpenSFS. *Introduction to Lustre* Architecture*, 10 2017. URL: http://wiki.lustre.org/images/6/64/LustreArchitecture-v4.pdf.

[4] OpenSFS. *Lustre* Software Release 2.x Operations Manual*. OpenSFS, 12 2017. URL: http://doc.lustre.org/lustre_manual.pdf.

[5] Wikipedia. Lustre (file system). URL: https://en.wikipedia.org/wiki/Lustre_(file_system) [cited 2017-12-4].